In28 Minutes

Generative AI Leader

Generative AI Leader - Getting Started

In 28 Minutes

- Artificial Intelligence is in its **third golden phase** :
  - **First Phase (1960s–1970s)**
    - Early successes
  - **Second Phase (2000–2020s)**
    - Breakthroughs due to:
      - New ML and deep learning algorithms
      - Access to massive datasets
      - Availability of powerful computing hardware (GPUs, TPUs)
  - **Third Phase (Now)**
    - Accelerated by Generative AI
    - AI creates new content:
      - text, code, images, audio, and more

Generative AI Leader - Getting Started - 2

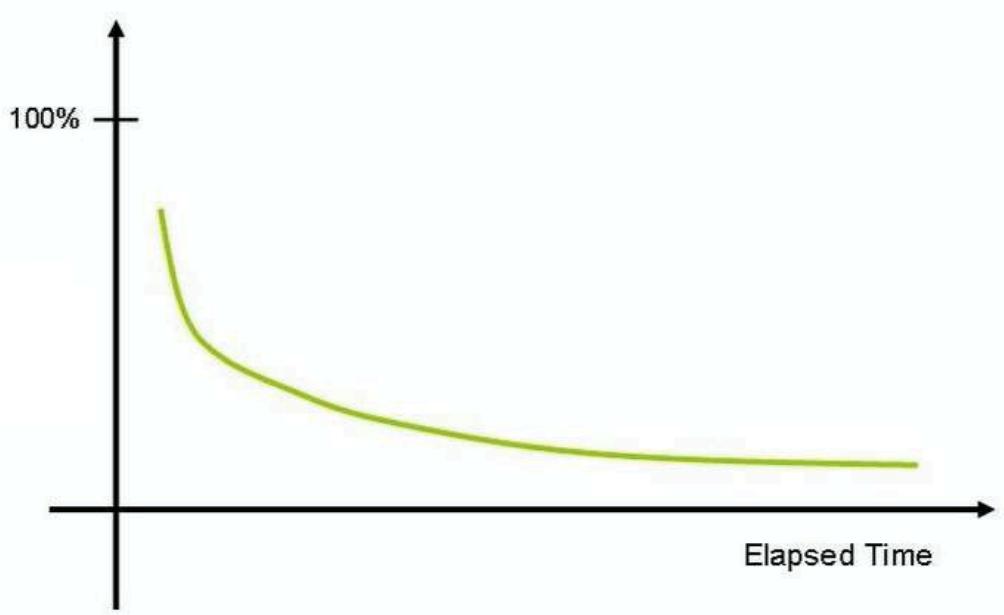
In 28 Minutes

- Building AI solutions has NOT been easy:
  - 🧠 **Skills Scarcity** : Limited availability of AI/ML talent
  - 🗃️ **Data Complexity**: Need to create, clean, and manage massive datasets
  - 🖥️ **Infrastructure Needs**: AI needs high performance computing, networking and storage
  - 🔄 **Model Lifecycle Management** : Complexities in training, tuning, deploying, and monitoring models
- How can we enable enterprises to **use AI** easily?
  - Google and Google Cloud offer a number of solutions
- **Our Goal** : Help you understand AI, ML and Generative AI while exploring various solutions offered by Google and Google Cloud (and help you get certified)

How do you put your best foot forward?

In28 Minutes

- **Tricky certification** - Expects you to understand and **REMEMBER** a number of new concepts and services
- As time passes, humans forget things.
- How do you improve your chances of remembering things?
  - **Active learning** - think and take notes
  - **Review** the presentation every once in a while



Our Approach

- Three-pronged approach to reinforce concepts:
  - 📽️ Presentations (Video)
  - 🎬 Demos (Video)
  - 🧪 **Two kinds of quizzes**:
    - 📝 Text quizzes
    - 📹 Video quizzes
- (Recommended) Take your time and do not hesitate to replay videos!
- (Recommended) Have Fun!

FASTEST ROADMAPS

in28minutes.com

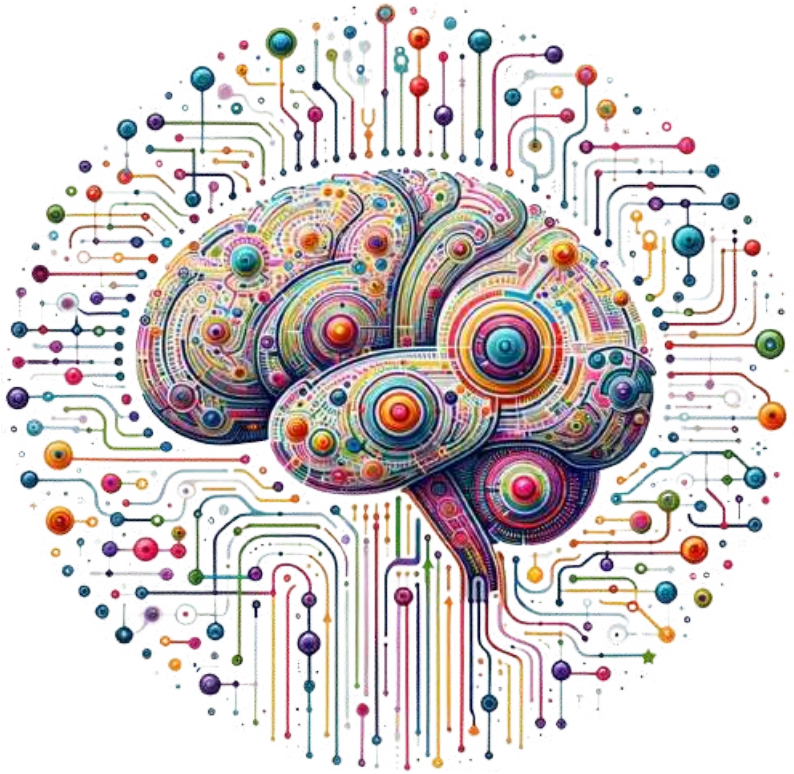
In28 Minutes

- Google Cloud Certifications
- Azure Certifications
- AWS Certifications
- DevOps
- Java Full Stack
- Java Microservices

Artificial Intelligence

Artificial Intelligence - All around you

- 🚗 **Self-driving cars**: Navigate roads and make real-time decisions
- 📩 **Spam Filters**: Keep unwanted emails out of your inbox
- 📨 **Email Grouping**: Automatically sort emails into categories
- 💳 **Fraud Detection**: Identify suspicious credit card transactions
- 🎯 **Recommendation Systems**: Movies, songs or product recommendations based on your preferences



What is AI? (Oxford Dictionary)

In 28 Minutes

The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

What is AI? (in Simple Terms)

- **Goal of AI**: Create machines that can simulate human-like intelligence and behavior
  - ♟️ Play Chess
  - 🛒 Make Purchase Decisions
  - 🚗 Drive a Car
  - 📝 Write an Essay!
  - 🎶 Compose Music
  - 🎨 Generate Art

Understanding Types of AI

In 28 Minutes

- **Strong AI (General AI)**: Machine intelligence = Human intelligence
  - A machine that can solve problems, learn, & plan for future
  - An expert at everything — even all sports and games!
  - Learns like a child, building on its own experiences
  - We are a little far away from this
    - Estimates: few decades to never
- **Narrow AI (Weak AI)**: Designed for a specific task
  - Example: 🚗 Self-driving car
  - Example: ♟️ Playing Chess
  - Example: 🏠 Predicting House Price

AI Fundamentals - Scenarios

In 28 Minutes

Scenario	Solution
Categorize: Building a computer system as intelligent as a human. An expert at everything (all sports and games!)	Strong AI
Categorize: Building a computer system that focuses on specific task (Self-driving cars, virtual assistants, object detection from images)	Narrow AI (or weak AI)

AI vs ML vs Generative AI

AI vs ML vs Generative AI

In 28 Minutes

- **Goal of AI**: Create machines that can simulate human-like intelligence and behavior
  - What is ML?
  - How does Generative AI fit in?
- Let's get started on a Journey!

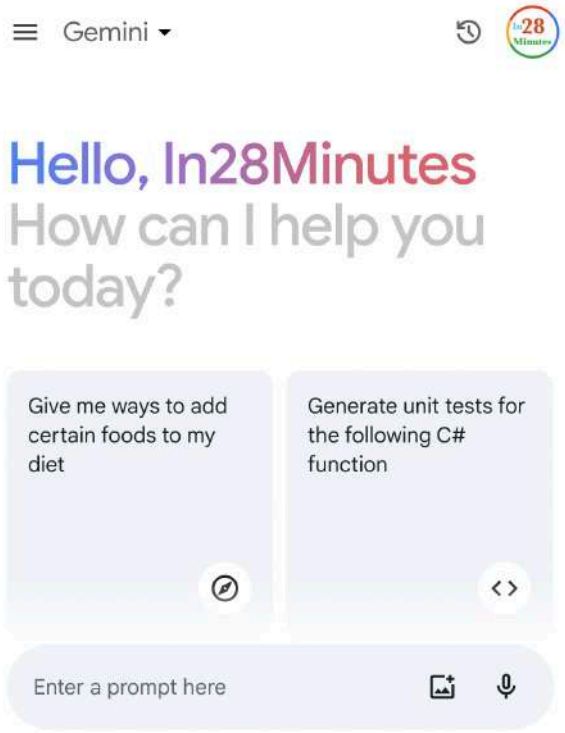
Playing with Gemini Chatbot

In 28 Minutes

- **Gemini**: Google's brand for its generative AI ecosystem
  - Models, chatbots, tools, and a lot more!
- **Let's play with Gemini - the chatbot:**
  - 💬 Can you make list of Top 10 technologies that I might want to learn as a cloud engineer?
  - 💬 Generate a bulleted list of items I need for an 15 day Everest Base Camp trek
    - 💬 I will be staying in tea houses on the trek. Can you update the list?
  - Gemini is multi-modal: Let's create an Image
    - 💬 Create an image showing the atmosphere at a cricket match
    - 💬 A vibrant, high-energy night scene at an Indian Premier League (IPL) cricket match in a packed stadium. Focus on the electric atmosphere: flashing LED boundary ropes, dynamic light shows, and the enthusiastic crowd, illuminated by floodlights. Show fans wearing team jerseys, waving flags, and cheering loudly. Capture the blurred motion of a batsman hitting a shot under pressure, with fielders poised. Include visible digital scoreboards and a general sense of excitement and celebration.

Playing with Gemini - Observations

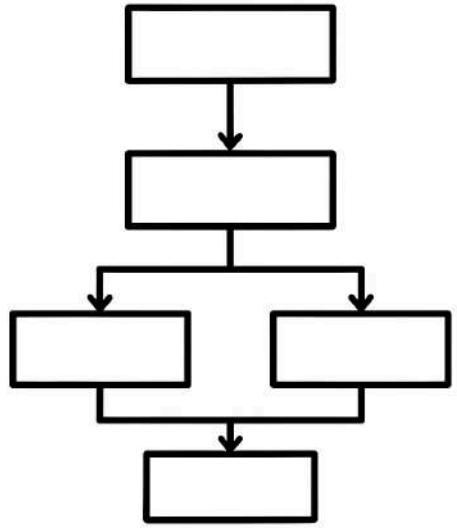
- Gemini MAY display inaccurate or offensive information
- BUT it does provide a lot of value if you understand its limitations and know when/how to use it
- How does Gemini work?
  - 🤖 Artificial Intelligence
  - 📈 Machine Learning
  - 🪄 Generative AI
  - 📚 Large Language Models
  - 🧱 Foundation Models


Gemini
Hello, In28Minutes
How can I help you today?
Give me ways to add certain foods to my diet
Generate unit tests for the following C# function
Enter a prompt here

Traditional Programming is based on Rules

In 28 Minutes

- **Traditional Programming**: Based on Rules
  - IF this THEN do that
- **Example**: Predict price of a home
  - Design an algorithm with predefined rules considering:
    - Location
    - Home size
    - Age
    - Condition
    - Market Trends
    - Economic indicators etc..



Machine Learning Learns From Examples

In 28 Minutes

- **Machine Learning**: Learns from examples (instead of rules)
  - **1:** Provide millions of examples
  - **2:** Train an algorithm to create a model
  - **3:** Use the model to make predictions on new data
  - **Example** - House Price Prediction
    - Data: Location, Size, Bedrooms, Age, Condition and Price of the home
    - ML learns the relationship from past sales data
  - **Example** - Handwriting Recognition
    - Data: Images of handwritten digits (0–9) along with the correct digit
    - ML learns to identify digits based on pixel patterns from the image data

Home size (Square Yds)	Age	Condition (1-10)	Price $$$
300	10	5	XYZ
200	15	9	ABC
250	1	10	DEF
150	2	34	GHI

Machine Learning Fundamentals - Scenarios

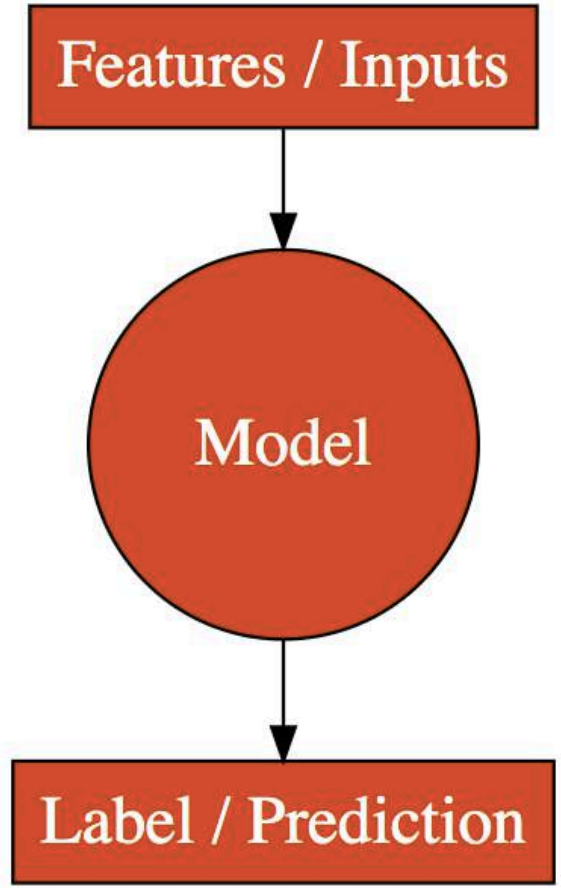
In 28 Minutes

Scenario	Solution
Category of AI that focuses on learning from data (examples)	Machine learning
How is ML different from traditional programming?	Traditional Programming: Rules. Machine Learning: Examples

Machine Learning - Making Prediction

In 28 Minutes

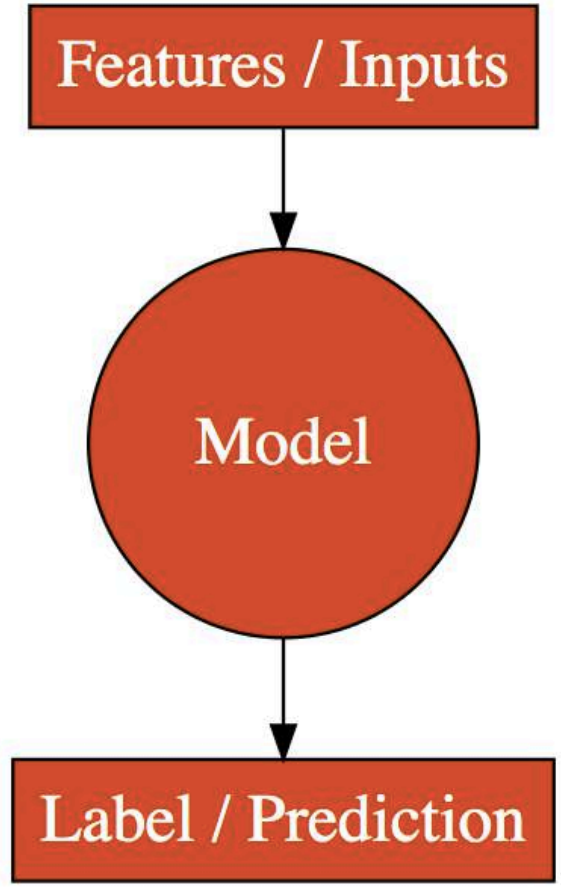
- **Goal**: Make a Good Prediction
  - Give inputs to a model
  - Model returns the prediction
  - Inputs are called **Features**
  - Prediction is called **Label**
  - **Example**: House Price Prediction Model
    - **Label**: price
    - **Features**:
      - area: Total area of house (m^2)
      - rooms: No. of rooms
      - bedrooms: No. of bedrooms
      - furniture: Is it furnished?
      - floor: Which floor?
      - age: How many years?
      - balcony: has balcony or not
      - garden: has garden or not



Machine Learning - Features and Labels - Examples

In 28 Minutes

- Used Car Price Prediction Model
  - **Label**: price
  - **Features**: manufacturer, year, model, age, condition, cylinders, location
- Spam Email Classification Model
  - **Label**: isSpam
  - **Features**: sender, subject, content
- Grant a Loan Model
  - **Label**: shouldWeGrantALoan
  - **Features**: doesOwnCar, doesOwnRealEstate, creditScore, isMarried, doesHaveChildren, totalIncome, totalCredit



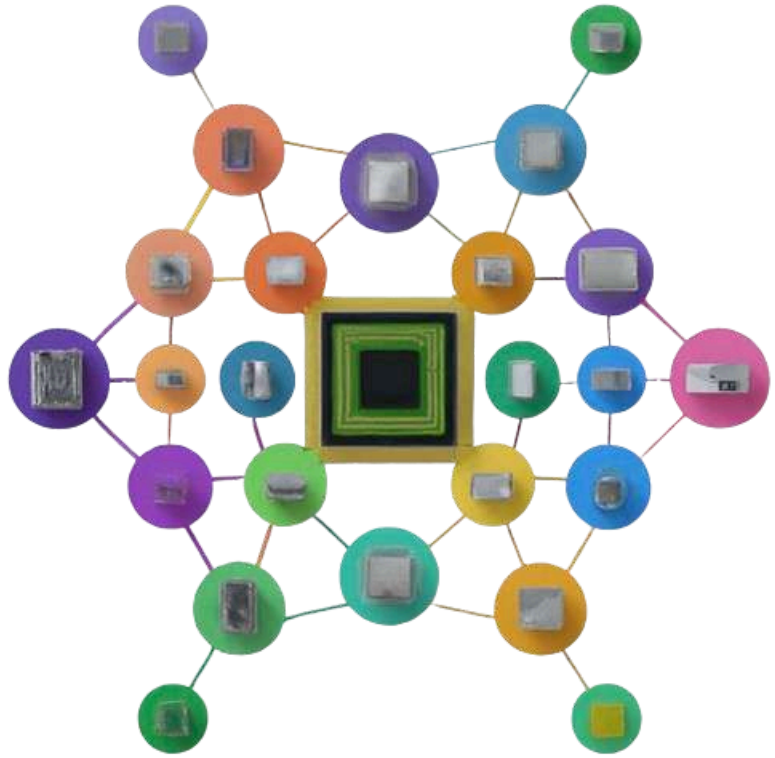
Machine Learning - Making Predictions - Scenarios

In 28 Minutes

Scenario	Solution
Categorize into features and labels for house price prediction: price, area, rooms, age	price is label. Others can be features
Categorize into features and label for used vehicle price prediction: manufacturer, year, model, age, condition, cylinders, location, price	price is label. Others can be features

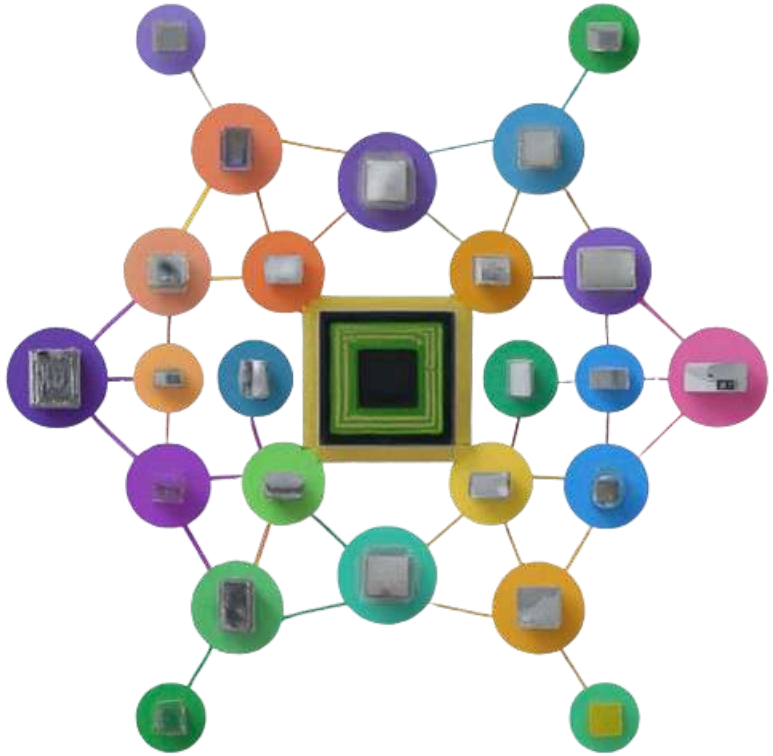
Creating Machine Learning Models - Steps

- Machine Learning uses a Step by Step Approach:
  - 📥 **1:** Data Ingestion
  - 🧹 **2:** Data Preparation
  - 🧠 **3:** Model Training
  - 🚀 **4:** Model Deployment
  - 📊 **5:** Model Management
- Let's now discuss each step - one at a time



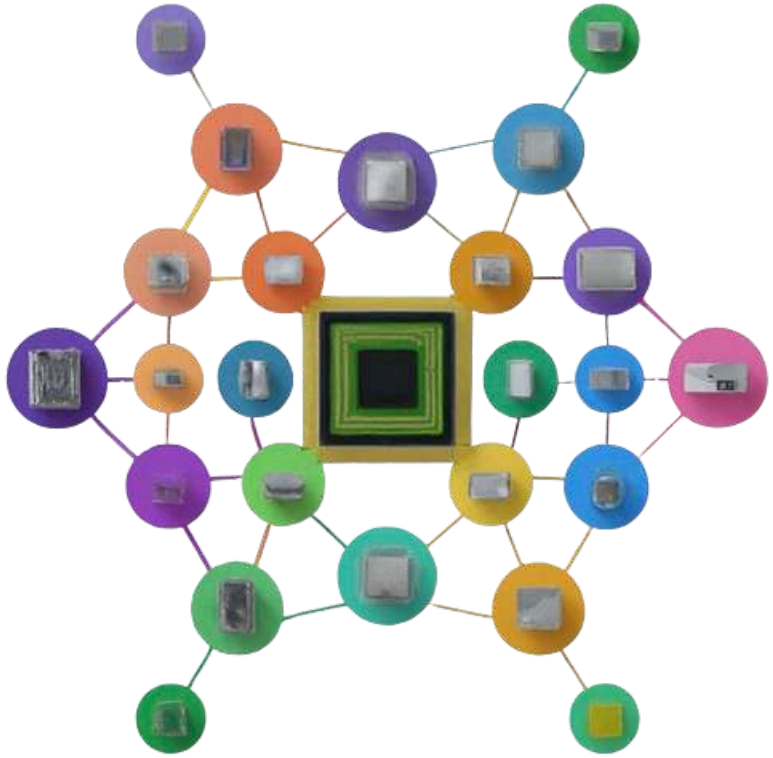
ML Lifecycle: Data Related Stages

- **Data Ingestion**: Collecting raw data from different sources
  - Streaming user activity from a mobile app
  - Pulling transaction logs from a point-of-sale system
- **Data Preparation**: Clean & format data for training
  - Removing duplicates from retail sales records
  - Handling missing values in patient health data



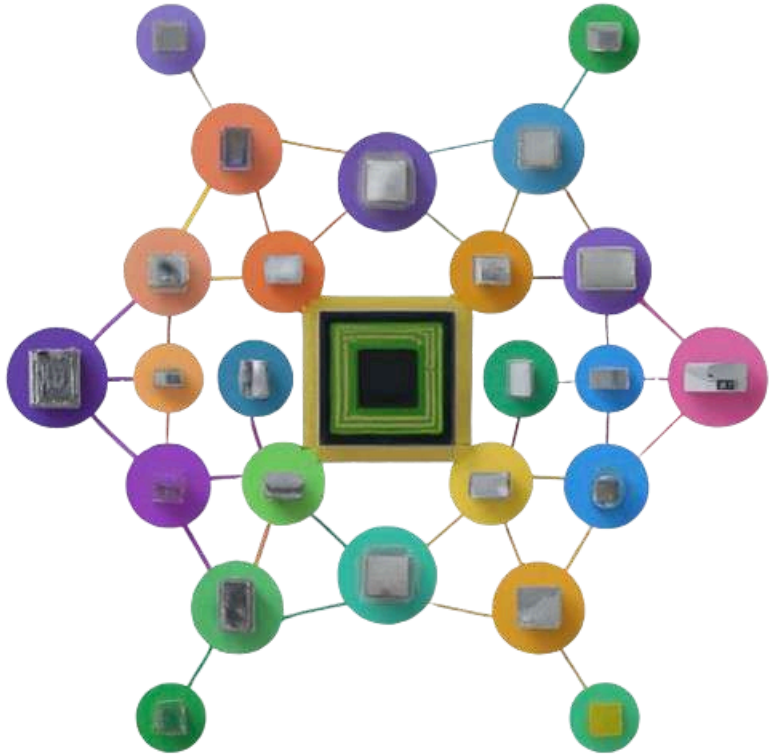
ML Lifecycle: Model Related Stages

- **Model Training**: Using data to teach your ML model
  - Training a fraud detection model using banking transactions
  - Training a price prediction model for used cars
- **Model Deployment**: Making the trained model available to users
  - Deploying a customer support chatbot on your website
- **Model Management**: Monitoring and maintaining your model post-deployment
  - Updating a model as new financial data comes in
  - Replacing a model if accuracy drops



ML Lifecycle Recap: End-to-End Example

- **Use Case**: ML model to detect spam in email
- **STEPS**:
  - **Ingest** your email history
  - **Prepare** by labeling spam vs not spam
  - **Train** with thousands of labeled examples
  - **Deploy** into your email platform
  - **Manage** performance, retrain as spam evolves



Identify the Stage in ML Lifecycle - 1

In 28 Minutes

Activity	Stage
Pulling product transaction logs from a retail store's log store	Data Ingestion
Streaming real-time user activity from a mobile app	Data Ingestion
Removing duplicate customer records from the training dataset	Data Preparation
Filling in missing values in patient data using averages	Data Preparation
Labeling emails as spam or not spam for training	Data Preparation
Using thousands of labeled housing records to teach a model to predict price	Model Training
Training a model to detect credit card fraud using past transaction data	Model Training

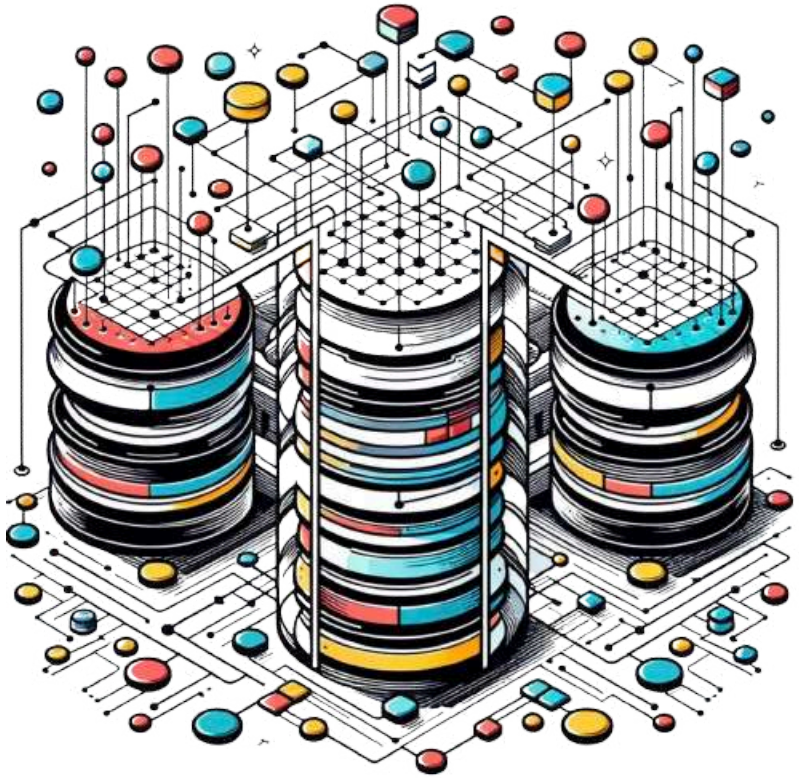
Identify the Stage in ML Lifecycle - 2

Activity	Stage
Making your trained model available via an API to a web application	Model Deployment
Deploying an image classification model into a mobile app	Model Deployment
Monitoring how accurate your chatbot is after deployment	Model Management
Updating your model with recent data after noticing performance drop	Model Management
Creating a dashboard that displays model inference accuracy trends	Model Management

Data is the Key!

Understanding Data Types

- **Structured Data**
  - Clearly organized and easily searchable
  - Examples:
    - Employee table - columns for ID, department, salary
    - Flight booking - fields for date, destination, seat number
    - Inventory table - SKU, stock count, and reorder level
- **Unstructured Data**
  - Free-form, lacks a predefined format
  - Examples:
    - Customer emails describing issues in their own words
    - User review on a website
    - A scanned handwritten letter from a client
    - A YouTube video recording of a product demo

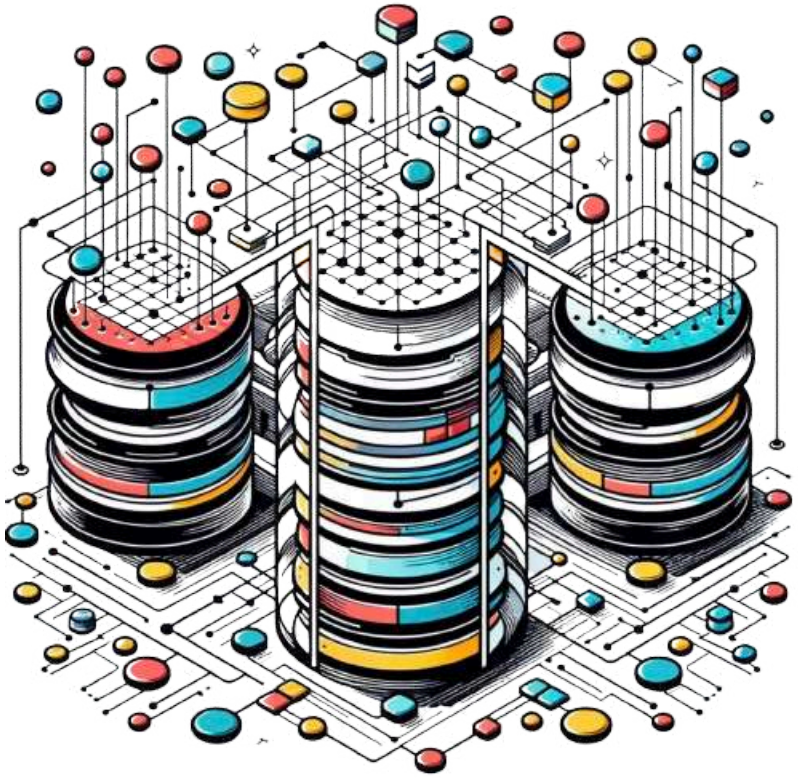


Identify Data Type

Scenario	Data Type
A dataset with columns for student name, grade, and attendance	Structured
A folder full of voice memos from customer support calls	Unstructured
An Excel sheet tracking orders by ID, product name, and delivery date	Structured
A Google Doc with brainstorming notes from a marketing meeting	Unstructured
A JSON file with customer profiles: name, email, and preferences	Structured
A scanned image of a signed contract	Unstructured

Understanding Labeled Data

- **Labeled data** = data tagged with meaning
  - Tags help ML models learn relationships
- Examples:
  - Images tagged as “cat” or “dog”
  - Customer reviews labeled: positive, negative, neutral
  - Emails marked as spam or not spam
- Used in **supervised learning**
  - The model learns by example and gets feedback
  - More labels = better learning, but labeling can be expensive



Understanding Unlabeled Data

- **Unlabeled data** = raw data with no predefined tags
  - Models must discover patterns without guidance
- Examples:
  - 🎵 Music files with no speaker ID or transcript
  - 🌐 Web traffic logs with no labeled source or action
  - 🖼️ A folder of random photos with no captions or tags
- Used in **unsupervised or semi-supervised learning**
  - Identify unusual patterns or outliers in data
    - Example: Spot error patterns in system logs
  - Group similar items together based on inherent characteristics
    - Example: Cluster music files by singer's voice signature
    - Example: Group photos based on who is in it

Identify Labeled or Unlabeled Data

In 28 Minutes

Scenario	Labeled or Unlabeled
You have a dataset of customer reviews marked as positive, neutral, or negative	Labeled
You're analyzing a folder of selfies, but none are labeled with names	Unlabeled
An e-commerce platform tags products as "electronics", "clothing", or "home"	Labeled
You receive thousands of support emails with no categorization	Unlabeled
A medical dataset marks X-ray images as "normal" or "abnormal"	Labeled
Network logs are collected but not tagged with user or activity type	Unlabeled

Supervised vs Unsupervised Learning

Supervised Learning: Learning from Labeled Data

- **Supervised learning** uses labeled data:
  - Input + correct output (or label)
- Model learns patterns and makes predictions on new data
- **Examples**:
  - Predicting loan approval based on income, credit score, and employment status
  - Classifying social media posts as "positive", "negative", or "neutral" sentiment
  - Identifying handwritten digits using labeled images from 0–9
- **Think**: “Learn from past → Predict the future”

Unsupervised Learning: Finding Hidden Patterns

- **Unsupervised learning** finds patterns in **unlabeled** data
- Clusters, organizes, or reduces data based on similarities
- **Examples**:
  - Organizing photo libraries by identifying similar faces or scenes
  - Grouping customers based on purchase history for targeted marketing
- Best for **discovery, grouping, and exploratory analysis**
  - Think: “Here’s a mess — can you make sense of it?”

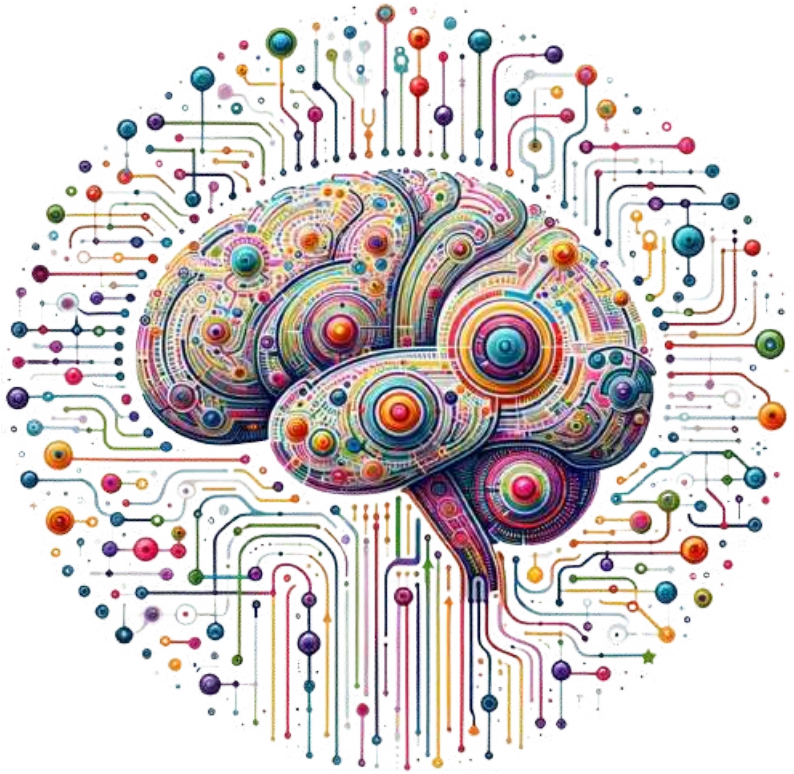
Supervised vs Unsupervised Learning

In 28 Minutes

Aspect	Supervised Learning	Unsupervised Learning
Data	Labeled	Unlabeled
Goal	Predict known outcomes	Discover hidden patterns
Examples	Email classification, fraud detection	Customer segmentation, anomaly detection
Output	Specific prediction or labels	Groupings, insights

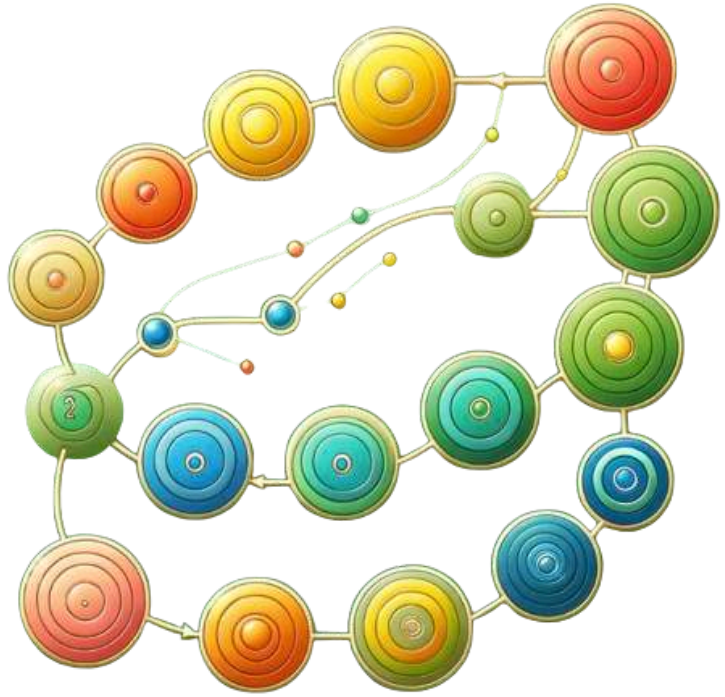
How Does a Child Learn to Walk?

- A child doesn't read instructions — **they learn by doing**
- They try walking, stumble, get up again, and adjust
- Feedback like applause, encouragement, or falling shapes their learning
- With every step, they learn balance, timing, and control
- This is the heart of reinforcement learning
  - **Try → Fail or Succeed → Learn → Try Again**



Reinforcement Learning: Learning by Trial and Error

- **Reinforcement learning** = learning by doing and getting feedback
- The model takes actions, gets rewards or penalties, and adjusts
  - **Examples**:
    - Game-playing AI learning to win chess by playing thousands of games
    - Short videos platforms changing video suggestions based on real-time feedback
    - A robot arm learns to pick up and place objects by experimenting with different motions and adjusting based on success or failure
- Ideal for **decision-making in dynamic environments**
  - Think: “Try → Learn → Improve”



Supervised vs Unsupervised vs Reinforcement Learning

In 28 Minutes

Scenario	Learning Type	Why?
Predicting house prices using features like area, location, and number of rooms	**Supervised**	We have labeled data with known prices to train the model.
Grouping customers into segments based on buying behavior	**Unsupervised**	There are no predefined labels; the model finds hidden patterns.
A robot learning to walk by trial and error	**Reinforcement**	The robot learns through feedback — rewards for walking, penalties for falling.
Email classification into "spam" and "not spam"	**Supervised**	The model is trained on emails that are already labeled.
A drone navigating through an obstacle course by learning from collisions	**Reinforcement**	Learns a strategy through repeated trials and feedback.
A game-playing AI learning how to win through experience	**Reinforcement**	The AI learns by receiving points or penalties based on actions taken.

THE AI Turmoil

In 28 Minutes

- **Quotes:**
  - I am really quite close, I am very close, to the cutting edge in AI and it scares the hell out of me - **Elon Musk**
  - The development of full artificial intelligence could spell the end of the human race. It would take off on its own, and re-design itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete and would be superseded. - **Stephen Hawking**
- No one knows the truth:
  - Most predictions about AI turned false in the last few decades!
- What's the pragmatic way to think?
  - Don't fear AI
  - Learn to make the best use of it

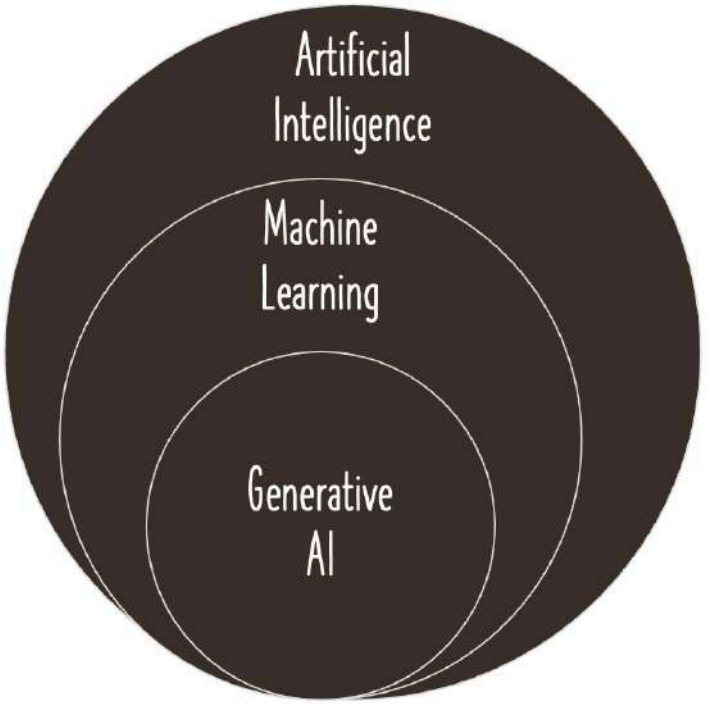
In28 Minutes

Generative AI
Foundation Models

Generative AI - How is it different?

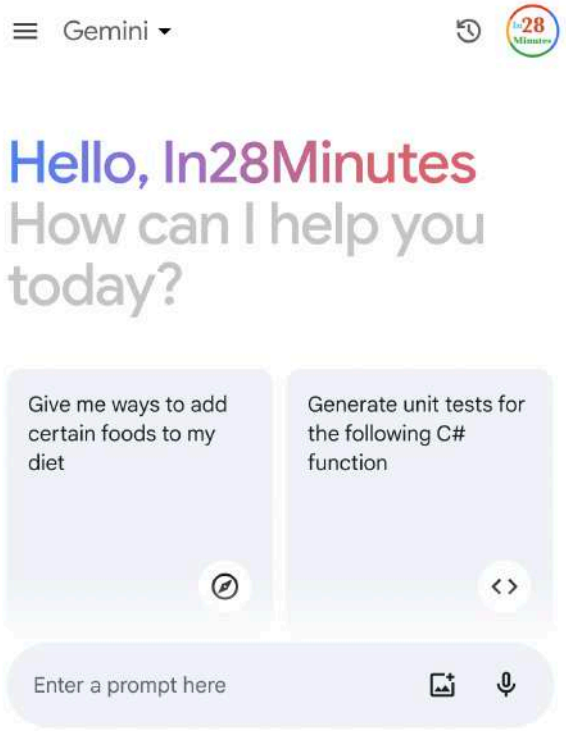
In 28 Minutes

- **Artificial intelligence (AI)**: Create machines that can simulate human-like intelligence and behavior
- **Machine learning (ML)**: A subset of AI where machines learn from data.
- **Generative AI**: An subset of ML that creates new content.



Generative AI - Generates New Content

- **Goal**: Generate New Content
  - Instead of making predictions, Generative AI focuses on creating new content
  - **Examples**:
    - **Text Generation**: Writing e-mails, essays & poems. Generating ideas. Generate responses for customer service interactions. Draft professional communication for clients.
    - **Code Generation**: Generate code snippets. Create unit tests.
    - **Media Creation**: Design images. Create instructional videos.
- How else is Generative AI different?
  - Let's find out!


Gemini
Hello, In28Minutes
How can I help you today?
Give me ways to add certain foods to my diet
Generate unit tests for the following C# function
Enter a prompt here

Generative AI uses Foundation Models

- Traditional models are **narrow** — trained for a specific task
  - 🎯 Example: A model that detects spam emails can't generate a marketing campaign
- What if we had models trained on **large, diverse datasets** across domains?
  - Enter **Foundation Models**
    - 🧠 Pretrained on **massive datasets**
      - REMEMBER: Needs complex training with huge infrastructure!
    - 🔄 Adaptable to a wide variety of tasks
    - 🧱 Foundation for many Generative AI apps

Foundation Models Have Complex Training Processes

In 28 Minutes

- **Complex Training Process**: Training is computationally intensive and complex
- **Huge Volumes of Data**: Requires learning from billions of text tokens or millions of labeled images
  - **Text Models**: Trained on Wikipedia, books, code, etc.
  - **Image Models**: Trained on datasets with image–caption pairs
- **Specialized Hardware**: Training needs huge volumes of specialized hardware - GPUs (Graphics Processing Units), TPUs (Tensor Processing Units) and Distributed Storage

Large Language Models: Text Foundation Models

- Trained on **huge volumes** of diverse text data
  - Books, articles, websites, code, and more
- Learn to **predict the next token** in a sequence
  - Example: "The cat sat on the ..."
  - Model gives probabilities:
    - "mat" (40%), "table" (20%), "chair" (20%), "moon" (10%)
  - It usually picks the most likely token — but we can tune settings to make it more creative
    - e.g., using the temperature parameter
- **Usecases**: Chat, summarization, translation, Q&A, ..

Diffusion Models: Turning Noise into Art

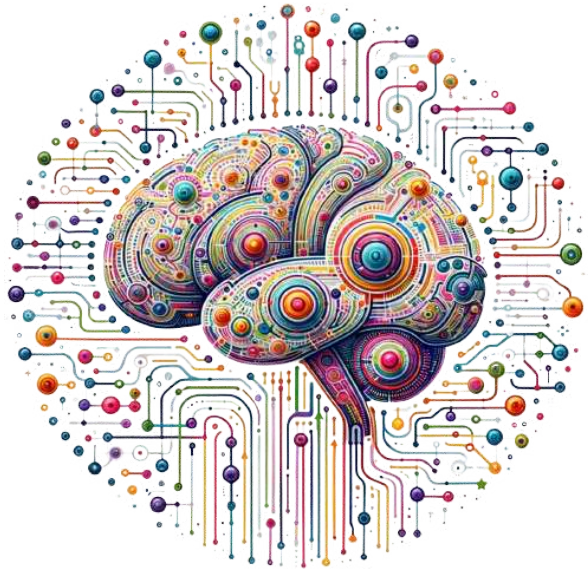
- **Diffusion models** are great at **generating images, audio, and video**
- **How they work**:
  - Start with **random noise**
  - Iteratively **refine it** step-by-step
  - Result: **Structured, realistic output**
- **Used for**:
  - Creating AI art and realistic pictures from text prompts
  - Generating lifelike human voices or music
  - Producing short AI-generated videos

Foundation Models - Scenarios

Scenario	Solution
Models trained on large, diverse datasets across domains. Adaptable to a wide variety of tasks.	Foundation Models
Text foundation models trained on huge volumes of diverse text data.	Large Language Models
Models that generate images, audio, or video starting with random noise and refining step-by-step	Diffusion Models

Foundation Model: Gemini

- Trained on: **Text, images, code, audio, video**
- **Multimodal** capabilities:
  - Understand and generate language
  - Write and explain code
  - Create images and videos
- **Enables tasks** like:
  - Answering complex questions from documents
  - Generating Python code from natural language prompts
  - Creating a new image from natural language prompts
  - Explaining how a diagram or chart works
  - Understand text, image and audio as part of a single task (Multimodal task)

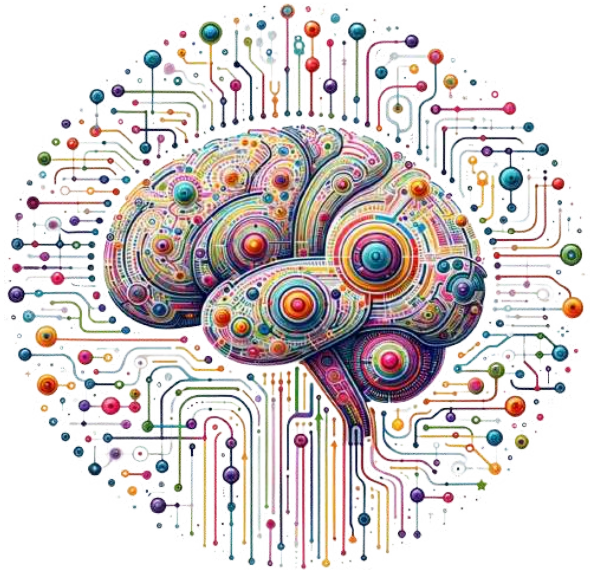


Foundation Model: Imagen

- Trained on: **Images + Text descriptions**
- **Specializes** in:
  - Text-to-image generation
  - Image editing from prompts
  - Understanding image content
- **Useful** for:
  - Designing social media graphics with text prompts
  - Generating illustrations for stories or articles

Foundation Model: Chirp

- **Trained on**: Massive multilingual audio datasets
- **Focused** on:
  - **Speech recognition**
  - **Voice-based interaction**
- **Useful** for:
  - Real-time transcription of meetings, calls, and lectures
  - Audio translation across languages
  - Voice assistants that understand user commands



Foundation Model: Veo

- **Foundation Model**: Google's video generation foundation model
  - "Transforms creative ideas into compelling video narratives using Google's advanced video generation model."
  - "Veo is capable of generating videos with audio from text prompts, or animating images with textual guidance."
- Use cases:
  - **Text-to-Video**: Generate dynamic video sequences directly from text prompts.
  - **Image-to-Video**: Transform static images into moving videos following text prompts.

Play with Foundation Models - Google AI Studio

In28 Minutes

- **Easy experimentation**:Ideal for learners and experimentation on a small scale
- **Needs Google Account**: Requires login with your Google account
- **Quick Prototyping**: Useful for prototyping ideas and designing effective prompts
- **Key Features**:
  - Easy-to-use, code-free interface
  - Built-in model playground with foundation models
  - Generate code to execute Gemini APIs
  - Build apps with Gemini

Foundation Model: Gemma

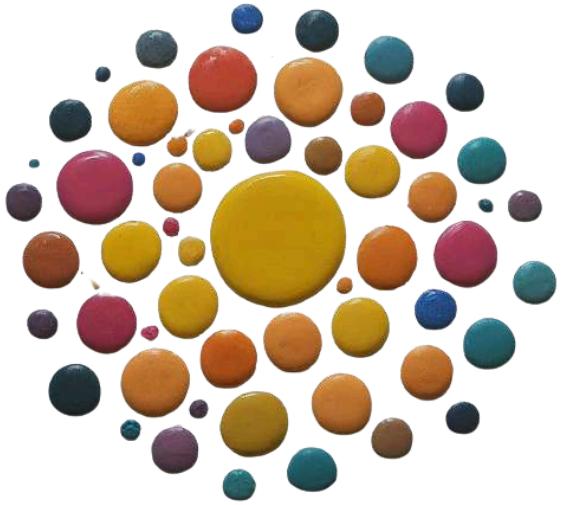
- **Gemma**: Family of lightweight, state-of-the-art open models
  - Built using the same research and technology behind Gemini
  - Openly released to help the AI community innovate and extend
- **Family of models**:
  - **Gemma**: Handles diverse generative AI tasks involving text and image input
  - **CodeGemma**: Specialized in programming tasks, lightweight and code-centric
  - **PaliGemma**: Build visual data processing AI solutions (Takes both images and text as inputs and can answer questions about images)

Exploring Foundation Model Limitations - 1

- **Data Dependency**: Output is only as good as the input
  - *Ex: A legal AI trained on outdated laws might recommend actions that are no longer valid*
- **Knowledge Cutoff**: Models don't know what happened after their training date
  - *Ex: A model trained in 2030 will not know about a product launched in 2040*
- **Hallucinations**: AI outputs that sound right — but are wrong
  - *Ex: An AI-generated academic citation pointing to a paper that doesn't exist in any journal*

Exploring Foundation Model Limitations - 2

- **Bias**: AI may reflect and amplify social, cultural, or historical biases
  - *Ex: A resume screening tool preferring male candidates due to skewed training data from past hiring*
- **Fairness**: Models may perform poorly on under-represented groups
  - *Ex: A facial recognition system with low accuracy for people with darker skin tones*
- **Edge Cases**: Rare or unusual situations expose model weaknesses
  - *Ex: A banking chatbot misclassifying a transaction because it doesn't recognize international wire transfer codes*



Reducing Hallucinations and Improving Accuracy

- **Grounding**: Connect AI to trusted data
  - *Ex: Connect AI to product documentation to answer customer queries accurately*
  - **Benefits**:
    - Reduces hallucinations
    - Provides citations
    - Improves trust and reliability
  - **Example Approach: Retrieval-Augmented Generation**:
    - Uses a search step to retrieve relevant info before generation
    - *Ex: RAG-enhanced chatbot that pulls from internal knowledge base to answer IT queries*
    - Steps:
      - Retrieve meaningful content
      - Add to prompt (augment)
      - Generate a grounded, accurate response

Foundation Model Limitations: Scenarios

In 28 Minutes

Scenario	Limitation	Recommended Solution
A health chatbot trained on 2015 articles misinforms users about current COVID-19 treatments	Knowledge Cutoff	Grounding using up-to-date medical documentation
A model trained on old resumes prefers male resumes	Bias	Fine-tuning with inclusive data
A climate model trained only on data from Europe fails to generalize well for Africa	Data Dependency	Fine-tuning with representative global data
An AI-generated article includes fake citations	Hallucination	Grounding via internal documentation
A model misses legal exceptions in rare court cases	Edge Case	Fine-tuning with domain-specific legal cases
A product chatbot invents features that don’t exist	Hallucination	Grounding with internal product docs and FAQs

In28 Minutes

Prompt Engineering

Prompting: Your Key to Great Gen AI Results

In 28 Minutes

- Prompting is the art of telling a Gen AI model **what you want**
- Better prompts = Better outcomes
- **Practice is essential** — every interaction is a chance to improve
- Techniques include:
  - **Examples** (zero-, one-, few-shot)
  - **Role prompting**
  - **Prompt chaining**
  - **Chain-of-thought prompting**
  - **ReAct prompting**

ZERO SHOT vs ONE SHOT - Example

In28 Minutes

- **EXAMPLE 1**: ZERO SHOT
  - Please choose the right answer:
  - Question: Which of these is a programming language?
    - A) Docker
    - B) Python
- **EXAMPLE 2**: ONE SHOT
  - Use the same answer format as the example.
  - Please choose the right answer:
    - Question: Which of these is a container orchestration tool?
      - A) Docker
      - B) Kubernetes
    - Answer: B is correct
    - Question: Which of these is a programming language?
      - A) Docker
      - B) Python
    - Answer:

ZERO SHOT vs ONE SHOT vs FEW SHOT - Example

In28 Minutes

- **EXAMPLES:**
  - **EXAMPLE 1**: ZERO SHOT
    - For the given order, return a JSON object
      - Order: A pizza and a pepsi
  - **EXAMPLE 2**: ONE SHOT
    - For the given order, return a JSON object
      - Order: A pizza and a pepsi
      - Output: {"pizza": 1, "pepsi": 1}
      - Order: A burger and a soda
      - Output:
  - **Example 3**: FEW SHOT
    - For the given order, return a JSON object
      - Order: A pizza and a pepsi
      - Output: {"pizza": 1, "pepsi": 1}
      - Order: A burger and 2 sodas
      - Output: {"burger": 1, "soda": 2}
      - Order: A burger, A pizza and 2 sodas
      - Output:

Role Prompting

- Assign the model a **persona or role** to set tone and style
- *Examples*:
  - *“You are a customer service agent helping a frustrated user”*
- Helps tailor tone, format, and language to match your needs
- Usually configured in "System Instructions"

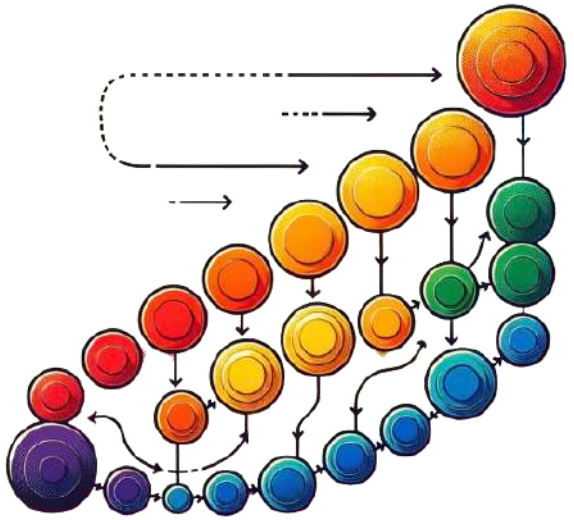
Need for Prompt Chaining

- You ask a foundation model a complex question:
  - Does it always respond **accurately and completely**?
    - **Not always**
- Does the model **struggle with multi-step reasoning** or **multi-part tasks**?
  - **Yes**, especially if the prompt is vague or overloaded
- Does breaking down the task help improve results?
  - **Absolutely!**
- How can you structure your prompts to guide the model better?
  - Use **Prompt Chaining / Step-by-Step Prompting**

What is Prompt Chaining?

In 28 Minutes

- A method to guide the model through **complex tasks in stages**
- Breaks a single large prompt into **smaller, linked prompts**
- Output of one prompt can be **input to the next**
- Helps the model stay **focused and accurate** at each step



Prompt Chaining: Step-by-Step Examples

In 28 Minutes

Goal	Prompt Chain
Draft a customer email campaign	1️⃣ Start with "Summarize our new product features" 2️⃣ Then prompt "Write a friendly email using that summary" 3️⃣ Final prompt: "Polish for a professional tone"
Plan a 3-day trip to Paris	1️⃣ "List top attractions in Paris" 2️⃣ "Group them into 3 daily itineraries" 3️⃣ "Add restaurant recommendations near each stop"
Prepare a job interview guide	1️⃣ "Summarize top skills for a Data Analyst" 2️⃣ "Generate 5 behavioral questions per skill" 3️⃣ "Add tips for answering each question"

Chain-of-Thought Prompting

- **CoT** = Guide the model to use step-by-step approach
  - **CoT Examples**:
    - “If a pencil costs ₹10 and a notebook costs ₹30, how much for 3 pencils and 2 notebooks? **Show steps.**”
    - “Solve: A train travels 60 km in 1.5 hours. What is its average speed? **Show your calculation.**”
    - ”Why does it rain more in coastal regions? **Explain step-by-step.**”
    - “Explain why binary search is faster than linear search, **step-by-step.**”
    - “This Python function isn't working. Can you debug it **step-by-step**?”
- Enhances **accuracy** and **explainability**

What is ReAct Prompting?

In 28 Minutes

- A prompting technique that alternates between:
  - **Reasoning**: Think through the problem logically
  - **Acting**: Perform actions (like a search)
- Enables models to **reflect**, take **intermediate steps**, and **improve answers**
  - Commonly used in **tool-augmented** workflows
  - Mimics how humans *think and do* iteratively

ReAct Prompting - Finance Example

- **User Question**: *"Is Apple a good investment right now based on its P/E ratio compared to the tech industry average?"*
  - **Model Reasoning Step**: *I need Apple's current P/E ratio*
    - **Model Action Step**: *Search: Apple current P/E ratio*
      - *Result: Apple's P/E ratio is 28.5*
  - **Model Reasoning Step**: *I need the tech industry's average P/E ratio*
    - **Model Action Step**: *Search: Tech industry average P/E ratio*
      - *Result: Industry average is around 25.0*
  - **Model Reasoning Step**: *Apple's P/E ratio is slightly higher than the industry average. This may indicate that it is slightly overvalued — or investors expect higher growth.*
  - **Model Final Answer**: *Apple's P/E ratio of 28.5 is above the tech industry average of 25.0, suggesting a premium valuation. Whether it's a good investment depends on your risk tolerance and growth outlook.*

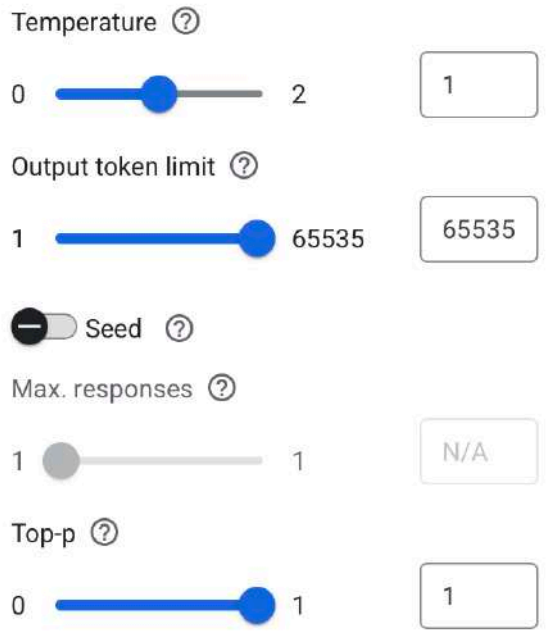
Prompting Techniques: A few scenarios

In 28 Minutes

Scenario	Best Technique	Why?
A customer service chatbot needs to handle complaints empathetically and provide refund information	**Role Prompting**	Assigning the model the role of a polite and helpful support agent ensures consistent tone and behavior
You're solving a math word problem and want the AI to show every step clearly	**Chain-of-Thought Prompting**	Encourages the model to think step-by-step and provide detailed reasoning
You're developing an interview bot that behaves like a career counselor giving personalized advice	**Role Prompting**	Setting the persona helps tailor tone and guidance for the scenario
A medical student asks the AI to diagnose a patient based on a case description with symptoms, history, and test results	**Chain-of-Thought Prompting**	The model needs to analyze symptoms step-by-step, consider possible diagnoses, rule out others, and explain the reasoning before arriving at the conclusion

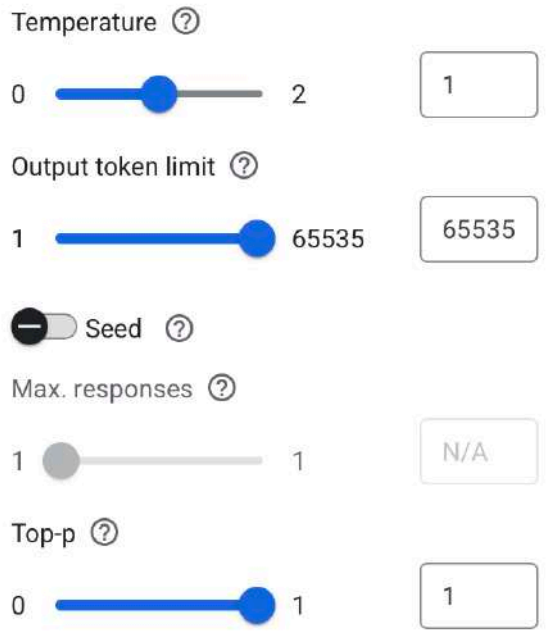
Experiment with Parameters

- **Output token limit**: How long do you want the response to be?
  - Specify maximum number of tokens in response
  - Remember: A token is approx. 4 characters. 100 tokens approximate to 60-80 words.
- Next set of parameters help you determine which token is chosen!
  - Example: A:40%, B:20%, C:10%, D:5%, E:2%, F:1%, ..
  - Which of A, B, C, D, E, F should be chosen?
  - Temperature, Top-P
  - This is an Art (NOT a science)



Experiment with Parameters - 2

- A:40%, B:20%, C:10%, D:5%, E:2%, F:1%, .. Which of A, B, C, D, E, F should be chosen?
- **Top-P**: What is the (cumulative) probability limit ?
  - Define the cumulative probability cutoff for selecting tokens
  - Lower value => less random responses. Higher value => more random responses.
    - Example: top_p value is 0.6 (60%) => Next token is either A or B
- **Temperature**: How random should be the output?
  - Higher values => more randomness (more creative)
  - Lower values => lesser randomness (more accurate)
  - Example Scenarios:
    - Find Capital City of India: use low value
    - Write a creative essay: use high value



Experiment with Parameters - Scenarios

In 28 Minutes

Scenario	Solution
Low or High Temperature: A student asks “What is 2 + 2?”	Low Temperature (You want an accurate and deterministic answer.)
Low or High Temperature: Writing a poem about "the color of silence"	High Temperature(Encourages imaginative and unexpected outputs.)
Low or High Top-P: Generating a formal email reply to a customer complaint	Low Top-P (Keeps output within a narrow, safe range)
Low or High Top-P: Brainstorming character names for a fantasy novel	High Top-P (Allows a wider sampling of less likely and more creative outputs.)
Output token limit: Answering "Summarize an essay in 50 words"	Small Output token limit (Short response expected)

Gemini Advanced

What is Edge Computing?

- **Edge computing** brings AI closer to where data is generated
- Runs AI on **local devices or edge servers**, not in the cloud
- Ideal for real-time decisions and low-latency applications
- *Example*: A self-driving car can’t wait for the cloud — it must respond instantly

Why Run AI on the Edge?

- **Speed**: Instant responses — no waiting for cloud round-trips
  - *Ex: A drone avoids obstacles in real time*
- **Privacy**: Keeps data local to the device
  - *Ex: Personal voice recordings stay on your phone*
- **Offline Functionality**: Works even without internet
  - *Ex: Translation app that works in airplane mode*

Gemini Nano: AI on Your Device

- **Gemini Nano** = A compact, efficient AI model built for edge devices
- **Benefits**:
  - Private: Data stays on your device
  - Fast: Near-instant responses
  - Offline: Works without network connection
- **Where it's used**:
  - Pixel Phones (Call Notes, Recorder summary)
  - Android OS (AI Edge SDK for developers)

Gemini for Google Cloud: Your Cloud AI Companion

- **Gemini** is integrated into the Google Cloud ecosystem
  - Use **natural language** to interact with different services
- Available in:
  - **Cloud Console**: Create or manage resources
  - **BigQuery**: Write, fix, and explain SQL queries
  - **Looker**: Explore and visualize data through plain English
  - **Security Command Center**: Understand threats and get mitigation guidance
  - ...

What is Google Workspace?

- A **cloud-based productivity suite** by Google
- Combines **communication**, **collaboration**, and **productivity** tools
- Includes apps like:
  - **Gmail**, **Google Docs**, **Google Sheets**, **Google Slides**
  - **Google Meet**, **Google Drive**, **Google Calendar**, and more
- Used by **millions of businesses and individuals** worldwide

Gemini for Google Workspace

In 28 Minutes

- Gemini is built into your favorite apps:
  - **Gmail**: Write, summarize, and reply to emails
  - **Docs**: Help writing content
  - **Sheets**: Generate formulas
  - **Slides**: Create presentations
  - **Meet**: Capture notes or summarize meetings
- **Boost productivity** across teams with smart, contextual AI help

Other Gemini Integrations

- **Google Vids**: AI-powered video creator and editor
  - Automatically generate storyboards, scripts, and visuals
- **AppSheet**: no-code platform (build apps without writing a line of code)
  - Use natural language to define workflows and logic
  - Turn spreadsheets and data into fully functional apps
- **Gemini Code Assist**: An AI coding assistant
- We already talked about:
  - **Gemini for Google Workspace**: AI help in Gmail, Docs, Sheets, and more
  - **Gemini for Google Cloud**: Assist with code, architecture, and cloud operations

Need for Larger Context Windows

- In traditional language models:
  - Can they remember earlier parts of **long conversations** or **big documents**?
    - **Not always**
  - Can they provide answers based on **entire textbooks or research papers**?
    - Often **NO**
- Smaller context windows limit what the model can "see" at once
- What if we want the model to **understand, summarize, and reason** across 100+ pages?
  - We need **larger context windows**

Introducing Gemini's Context Window

In 28 Minutes

- Gemini supports a **context window of up to 1 million tokens**
- Enables processing of **entire books, codebases, documents, or hour-long videos**
- Gemini don’t forget what was said in the beginning!

Why Are AI Safety Settings Important?

In 28 Minutes

- Can AI models sometimes produce **harmful or biased responses**?
  - **Yes**
- Can they be used to generate **misleading or unsafe content**?
  - **Yes**
- As AI becomes more powerful, we need guardrails to:
  - Protect users
  - Ensure models act responsibly
  - Prevent misuse and harmful outcomes
- Enter **Gemini Safety Settings**

Introduction to Gemini Safety Settings

In28 Minutes

- Gemini models come with **customizable safety settings**
- Safety settings include:
  - **Harassment**
  - **Hate**
  - **Dangerous Content**
  - ...
- Safety settings can be adjusted to avoid harmful content

Build custom experts with Gems

In 28 Minutes

- **Gems**: Custom AI experts for help on any topic
- **Premade Gems**: Gemini offers a set of premade Gems
  - A career guide gem
  - A brainstormer gem
  - A coding partner gem
- **Build Your Own**: You can build your own gems as well!
  - **Getting Started**: Define context and configure your prompt
  - **Upload your own files**: Provide necessary context
  - **Personalize your experience**: Configure specific tone and style
- **Operate more efficiently**: Save highly detailed prompt instructions for your most repeatable tasks

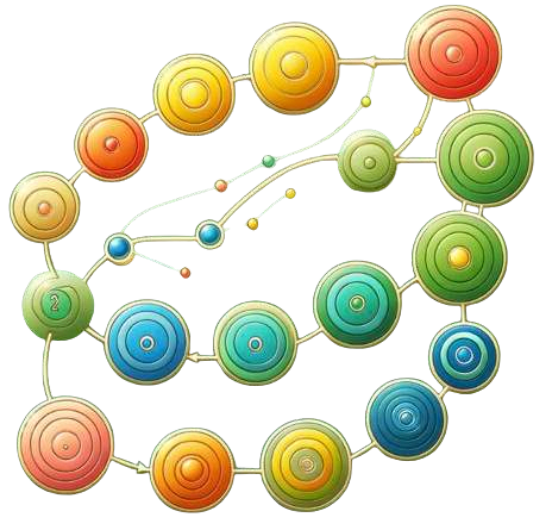
NotebookLM

What is Grounding?

- **Grounding** = Connecting an AI's output to **verifiable sources**
- Enhances **trust**, **accuracy**, and **relevance**
- In Gemini:
  - Add links, documents, or guidelines in your prompt
  - Gemini uses them to generate informed and tailored results
- *Example*: Using your company training documentation to help answer trainee questions

RAG: Expanding Model Capabilities

- **Before RAG**: LLMs could only use static training data
- **RAG = Retrieval + Augmentation + Generation**
- Empowers LLMs to access **external knowledge** in real-time
- Improves accuracy, relevance, and transparency

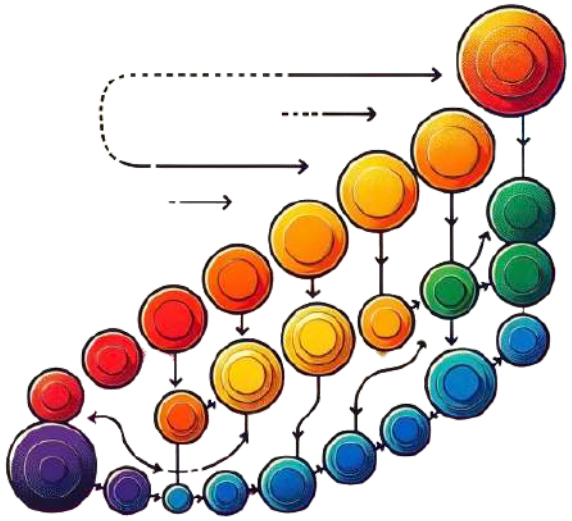


How RAG Works?

- **Retrieval**:
  - Search tools fetch context from data stores, APIs, or vector DBs
- **Augmentation**:
  - Retrieved info is added to the prompt
- **Generation**:
  - LLM responds using both query + retrieved context
- **Iteration (Optional)**:
  - Model refines its search if context isn't useful

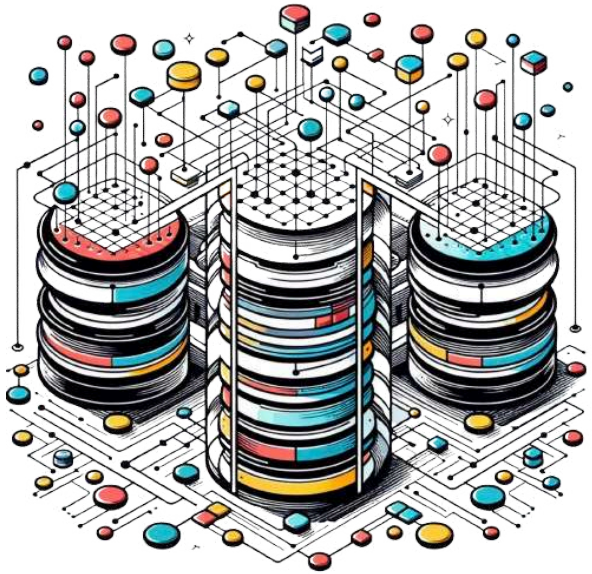
Why Use RAG?

- **Improves Accuracy**: Grounds answers in real, verifiable documents
- **Reduces Hallucinations**: Constrains responses to authorized knowledge only
- **Increases Transparency**: Supports source citations
- **Constrain LLMs**: Constrain model to only respond with *your* knowledge base
  - **Example**: Only answer using internal policy documents



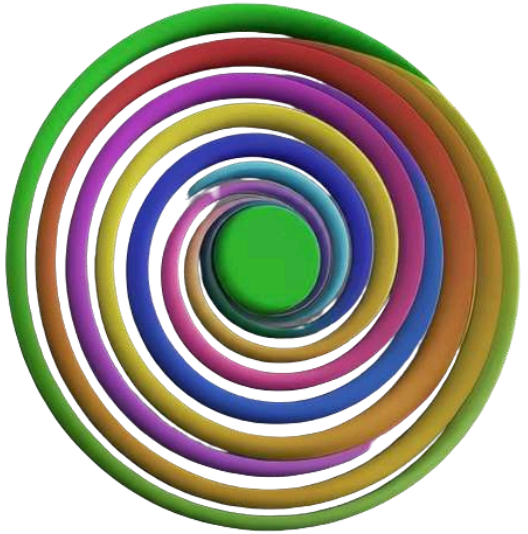
What Is NotebookLM?

- **NotebookLM** = AI-first research notebook using Gemini models
- Grounded in documents **you provide**
- Great for:
  - Research
  - Summarizing PDFs or presentations
  - Creating outlines, Q&A, speaker notes, or training materials
- Output always based on your uploaded sources — not random web data



How NotebookLM Works

- Upload docs (PDFs, slides, audio)
- Features:
  - Create one-click summaries, FAQs, and briefing docs
  - Ask questions for deeper insights and get answers with citations
  - Generate Audio Overviews and listen on-the-go
  - Create a Podcast Episode!



NotebookLM: Scenarios

Scenario	How NotebookLM Helps
A new hire needs to understand company policy documents	Upload the employee handbook and generate a summarized FAQ
A product manager needs to brief executives on a recent market report	Upload the report and generate a concise one-click briefing doc
A professor wants to prepare study guide for a long lecture slide deck	Upload the slide deck and generate study guide
A finance team wants a voice summary of quarterly results	Generate an audio overview from uploaded earnings reports

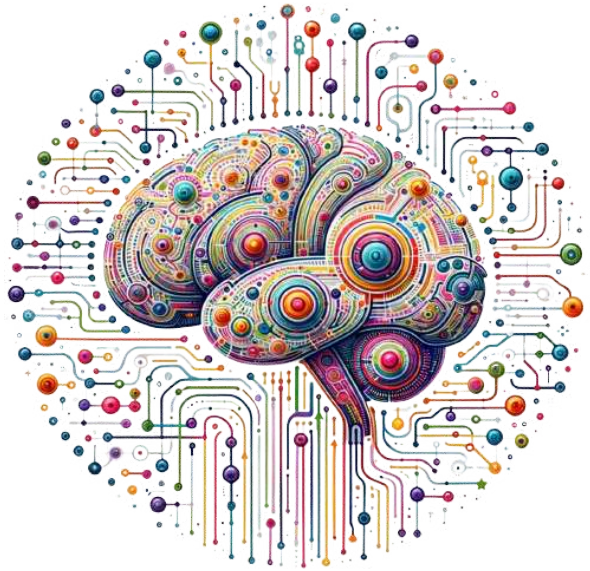
Google Cloud

Google's AI-First Journey - 1

- Sundar Pichai announced Google's AI-first strategy (2016)
  - Formal shift from mobile-first to AI-first
- Google has been using AI long before 2016:
  - **Google Search**: AI drives context understanding, ranking, and spelling correction
  - **Google Translate**: One of the first big tools to use AI for translating between different languages.
  - **Voice Search**: Google added deep learning to understand what you say in real-time.
  - **TensorFlow**: Google's powerful open-source ML framework

Google’s AI-First Journey - 2

- **Continued Innovation**: Google continues to drive AI innovations
  - **TPUs**: Custom hardware for fast and efficient model training
  - **Transformer Architecture**: Foundation of modern generative AI (used in Gemini, ChatGPT)
  - **Gemini**: Google’s multimodal foundation model
    - Integrated into several Google products: **Search**, **Workspace**, **Android**, and **Cloud**
  - **Responsible AI**:
    - Research into fairness, transparency, and safety
    - **SAIF (Secure AI Framework)** for secure and ethical AI deployments



What is Google Cloud?

In 28 Minutes

- Can every organization build world-class AI and infrastructure like Google?
  - **Not practically**
- Google launched **Google Cloud in 2008** to make its innovations available widely
- Google Cloud offers 200+ Services (compute, networking, storage, databases, AI/ML, ..)
  - Google Cloud simplifies AI and Gen AI: — So you don't have to build everything from scratch
  - **Vertex AI**: Unified platform for AI/ML
    - Build, deploy, and manage custom ML models
    - Make use of and fine tune Generative AI models

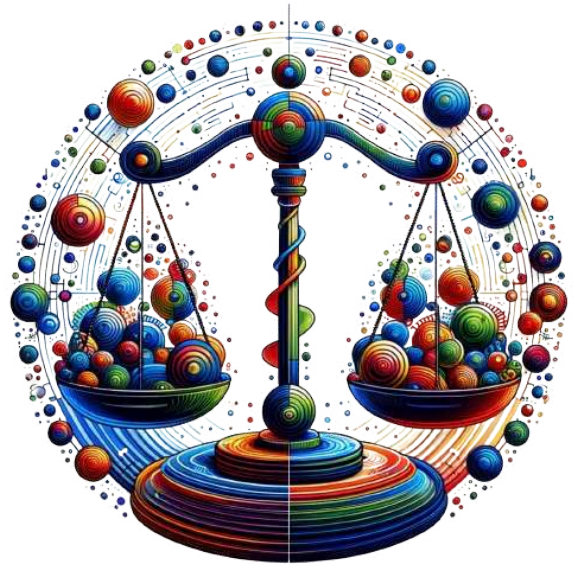
Google Cloud: Powerful Infrastructure For AI

In 28 Minutes

- **AI needs Great Infrastructure**: Training and Running AI Models Efficiently needs Great Infrastructure
- **Compute**:
  - **TPUs**: Custom Google hardware purpose-built for ML
  - **GPUs**: Great for parallel tasks. Ideal for Deep Learning.
  - **Hypercomputers**: Supercomputer-scale clusters of GPUs/TPUs
    - Unlock massive-scale training across distributed systems
- **AI-Optimized Storage**: High volume and High throughput storage for rapid access to large datasets
- **Global Fiber Network**: High bandwidth and Low latency

Google Cloud: Enterprise-Ready

- **End-to-End Security**
  - Includes data encryption at rest and in transit, IAM, and zero-trust networking
- **Comprehensive Compliance**
  - Certified for
    - ISO 27001
    - SOC 2
    - HIPAA
    - GDPR
    - and more



Why Google Cloud? - 1

- **Stay Business-Focused**: Apply AI to real problems without managing infrastructure
- **Leverage Existing Expertise**: Use Google's AI advancements without being an AI-first company
- **Access Advanced Technology**: Benefit from Google's cutting-edge models and tools
- **Scalable & Reliable**: Infrastructure ready to grow with your business
- **Accelerated Development**: Pre-built tools reduce time-to-market — no need to reinvent the wheel

Why Google Cloud? - 2

- **Open & Flexible**: Avoid vendor lock-in with support for open standards (TensorFlow, PyTorch, Kubernetes ..)
- **Continuous Improvement**: Automatic updates, security patches, and model upgrades
- **Guidance on Responsible AI**: Frameworks and tools for ethical, secure AI deployments
- **Get Enterprise-Grade Security & Compliance**: Trusted by industries with sensitive workloads

Scenarios: Why Google Cloud?

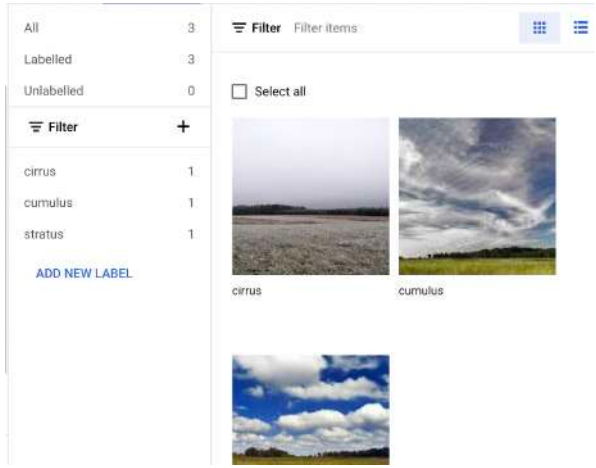
Scenario	Google Cloud Advantage
A fintech startup is building across clouds and wants to avoid being locked into a single vendor	**Open & Flexible** platform supports open-source frameworks and multi-cloud strategies
A growing e-commerce platform faces seasonal traffic spikes	**Scalable** infrastructure ensures performance under load
A media firm needs cutting-edge text-to-image and language models for content generation	**Access Foundation Models** like Gemini, Imagen,.. without building from scratch
A global manufacturing firm needs to ensure AI systems are fair and explainable	**Responsible AI Guidance** — with tools like Vertex Explainable AI and SAIF (Secure AI Framework)

ML in Google Cloud Before Generative AI

ML in Google Cloud - Traditional Landscape

In 28 Minutes

- This is the ML landscape before emergence of Generative AI
- **Machine Learning based API**
  - Natural Language, Vision, Speech etc
- **Custom Models** without needing ML expertise
  - Vertex AI > Auto ML
- Build **Complex Custom Models**
  - Vertex AI > Custom Training



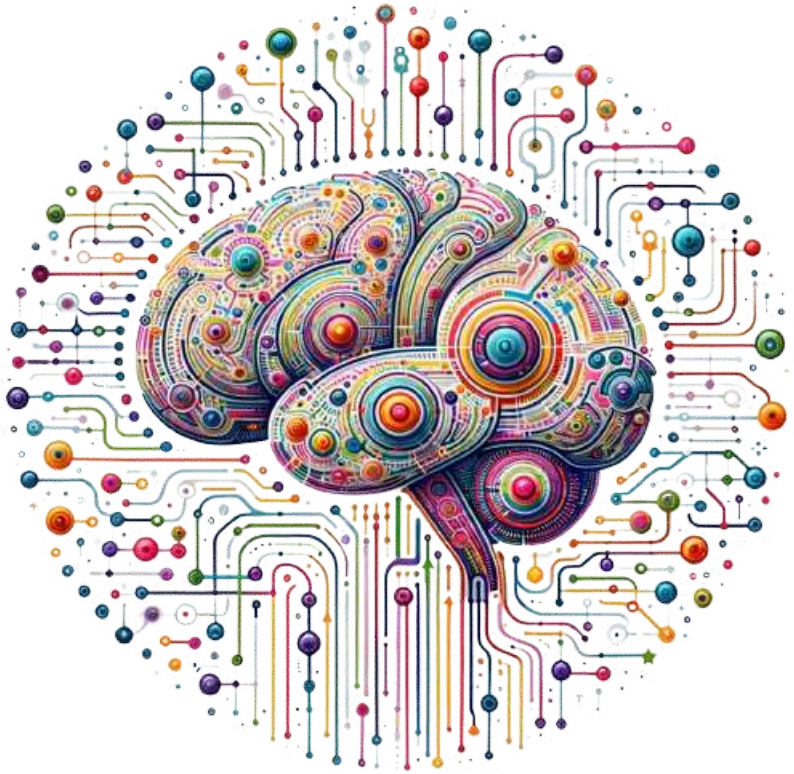
Google Cloud - 1 - Using Machine Learning API

In 28 Minutes

- **Usecase**: Derive insights from unstructured text
  - Natural Language API - ***https://cloud.google.com/natural-language***
- **Usecase**: Speech to Text
  - Speech to Text API - ***https://cloud.google.com/speech-to-text***
- **Usecase**: Convert text into speech
  - Text to Speech API - ***https://cloud.google.com/text-to-speech***
- **Usecase**: Extract insights from images, documents, and videos
  - Vision API - ***https://cloud.google.com/vision***
- **Usecase**: Content moderation for Video, Object detection and tracking
  - ***https://cloud.google.com/video-intelligence***

Google Cloud - 2 - Custom Models without ML expertise

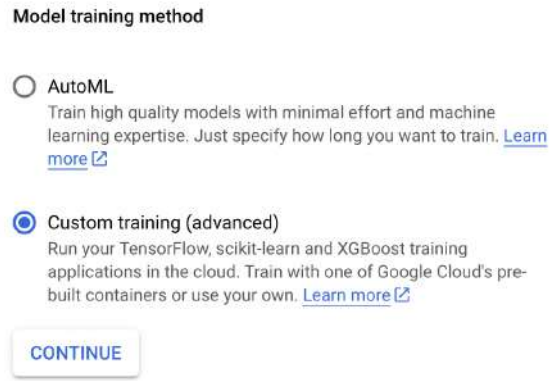
- What if your team **DOES NOT** have ML expertise but you want to build custom machine learning models?
- **Solution**: Vertex AI **AutoML**
- You want to build a custom image classification solution without ML expertise!
- **Example**: Identify the specific type of cloud
- **1:** Provide examples - Example images and categorization
- **2:** AutoML creates the model for you!



AI in Google Cloud - 3 - Build Complex Custom Models

In 28 Minutes

- You have a complex ML problem to solve
- You have a team with the skills needed
  - Data Scientists, ..
- You want to make use of ML Frameworks
  - TensorFlow
  - PyTorch
  - scikit-learn
  - ...
- **Solution: Vertex AI Custom Training**


Model training method
AutoML
Train high quality models with minimal effort and machine learning expertise. Just specify how long you want to train. Learn more
Custom training (advanced)
Run your TensorFlow, scikit-learn and XGBoost training applications in the cloud. Train with one of Google Cloud's pre-built containers or use your own. Learn more
CONTINUE

ML in Google Cloud After Generative AI

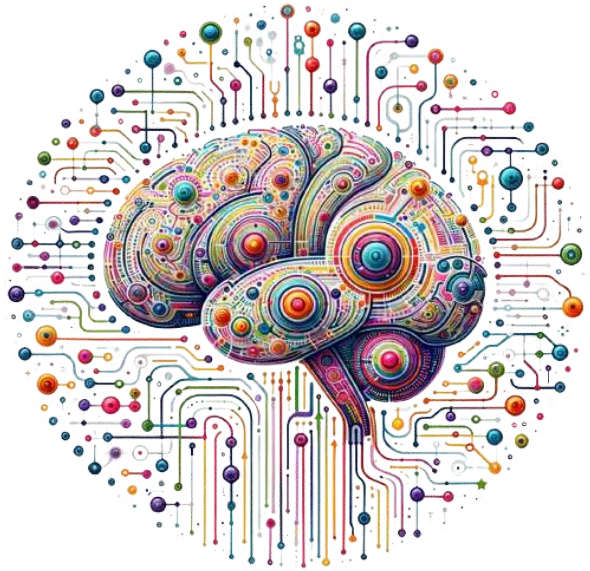
ML in Google Cloud – After Generative AI

- Generative AI brought **new capabilities**:
  - 🧠 Foundation models trained on vast, diverse datasets
  - 💡 Generate text, images, code, and audio
- Google Cloud has adapted by:
  - **Upgraded Machine Learning API** with generative capabilities
    - e.g., Natural Language API now includes summarization and content generation
  - **Vertex AI Studio**: Test, tune and deploy enterprise-ready generative AI
    - **Vertex AI Model Garden**: Access and finetune Foundation Models: Gemini, Imagen, and other models
  - **Vertex AI Agent Builder**: Build multi-step AI workflows using Agents

Flexible ML Pathways in Google Cloud - Basic

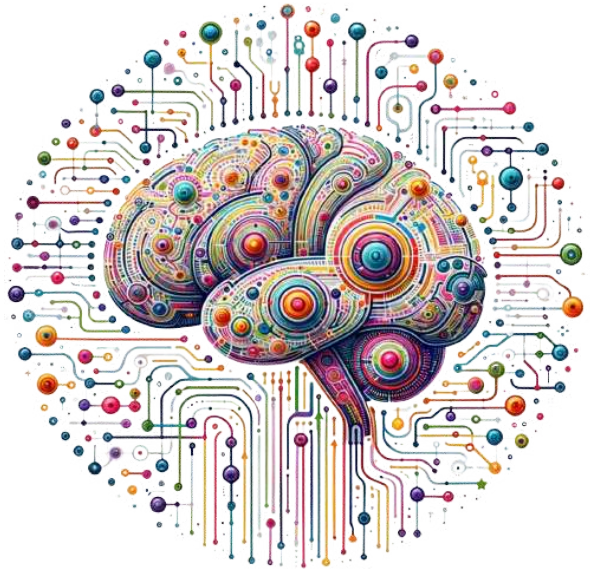
In 28 Minutes

- **1. Use Pretrained APIs**
  - ✅ Natural Language API, Vision API, Speech-to-Text, etc.
  - 🚀 Fastest way to integrate AI into applications — just call the API
- **2. Use Foundation Models (as-is via API)**
  - 🌐 **Vertex AI Model Garden**: Experiment with first-party models like **Gemini**, or Partner models like Llama
- **3. Customize Foundation Models**
  - 🎯 **Vertex AI Model Garden** + Fine Tuning: Fine-tune base models using your **domain-specific data**
  - 🗣️ Adjust tone, behavior, or style (e.g., legal, support)



Flexible ML Pathways in Google Cloud - Advanced

- **4. Train Your Own Models with AutoML**
  - 🚀 High-quality models without extensive ML expertise
  - 👩‍💻 For teams without deep ML expertise
    - BUT you want to build your own custom model from scratch!
  - ⚙️ Automatically handles all machine learning complexity
- **5. Build Complex Custom Models**
  - 🧪 You want full control over model creation
  - 👩‍💻 For teams with deep ML expertise
  - 🧰 Use frameworks like **TensorFlow**, **PyTorch**,..
  - 🚀 Solution: **Vertex AI Custom Training**



Flexible ML Pathways in Google Cloud – Scenarios 1

In 28 Minutes

Scenario	Solution
A company wants to build a chatbot using a foundation model without any fine-tuning	Use Foundation Models from Model Garden
A healthcare company wants their AI assistant to respond in a medical tone with accurate terminology	Fine Tune Foundation Model
An e-commerce platform wants an image generator that matches their brand aesthetic	Fine Tune a Foundation Model (with brand-specific data)
A support team wants to use a model that knows the company's internal tone and product line	Fine Tune Foundation Model (using support transcripts)

Flexible ML Pathways in Google Cloud – Scenarios 2

In 28 Minutes

Scenario	Solution
A retail team wants to build a model to categorize images but doesn't have ML engineers	Vertex AI AutoML
A telecom firm has a team of data scientists building a complex churn prediction model using TensorFlow	Build Complex Custom Models (Vertex AI Custom Training)
A logistics company wants to train a route optimization model using PyTorch	Build Complex Custom Models (Vertex AI Custom Training)
A bank wants to build and train a fraud detection model on GPU-based infrastructure with full control over training logic	Build Complex Custom Models (Vertex AI Custom Training)

Vertex AI

Vertex AI: Google Cloud's AI Platform

In28 Minutes

- Build, deploy, and scale AI — **all in one place**
- **Combines** infrastructure, tools, and API for:
  - Traditional ML Lifecycle AND
  - Generative AI usage and tuning

Vertex AI – Build Your Models From Zero

In 28 Minutes

- **Two Approaches To Build Models From Zero**
  - **AutoML**: Build high-quality models without deep ML expertise (Just provide the dataset)
  - **Custom Training**: Create and train models at scale using any ML framework - TensorFlow, PyTorch,.. (Take complete control)
- **End-to-end MLOps**: From data to predictions
  - **Fully managed infrastructure**: Scale infrastructure on demand for training and deployment
  - **Datasets**: Manage your training data
  - **Experiments**: Track and compare your ML experiments
  - **Model Registry**: Maintain model versions with complete tracking
  - **Model Monitoring**: Monitor your model; auto-trigger retraining if needed

Vertex AI Studio - Gen AI Made Easy

In 28 Minutes

- **Vertex AI Studio**: Rapid prototyping and testing of generative AI models
  - **Model Garden**: Hundreds of models from Google, partners, and open-source
    - *First Party Google Models*: Gemini, Imagen, Veo, ..
    - *Open Models*: Gemma (Lightweight, state-of-the-art open models from Google), CodeGemma, PaliGemma, Llama, Mistral
    - *Partner Models*: Claude (Anthropic), and more
  - **Prompt Gallery**: Explore ready-to-use prompts for common use cases
  - **Tuning**: Adapt foundation models to your domain with custom data

Google AI Studio vs Vertex AI Studio

In28 Minutes

Feature	Google AI Studio	Vertex AI Studio
Audience	Beginners and Experimenters	Enterprise teams
Access	Google account	Google Cloud
Use Case	Try prompts and models quickly	Build, tune, and deploy production-grade models
Governance & Security	Basic	Enterprise-grade access control and compliance

Agents

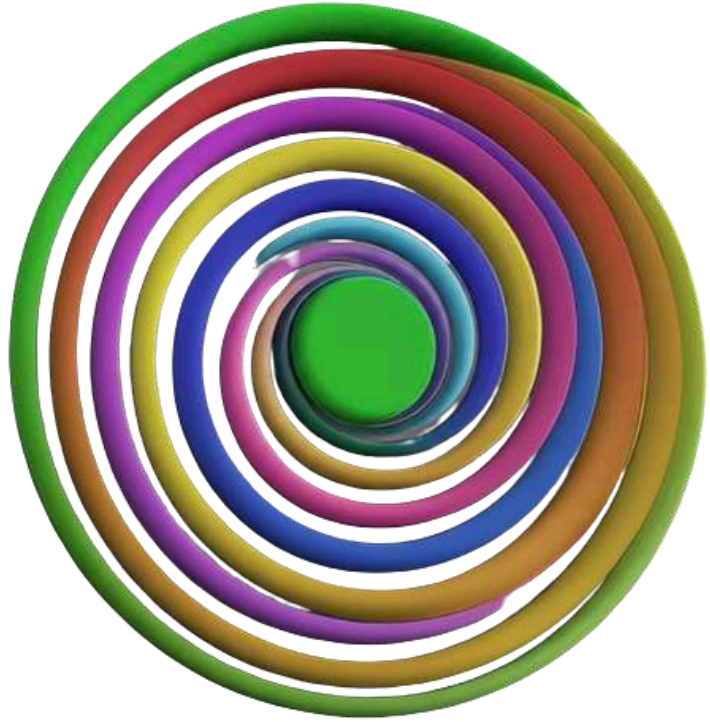
Why Do We Need AI Agents?

- Can a model act on its own to complete tasks?
  - **No** — It needs something to guide and execute
- Can you plan or make decisions using just a model?
  - **Not really** — Models are just one part of the solution
- So how do you build intelligent systems that **think and act**?
  - We need an **AI Agent**

What Are AI Agents?

In 28 Minutes

- **AI Agents** are systems that **observe**, **reason**, **decide**, and **act**
- Think of them as **goal-driven assistants**
- They don't just answer — *they take action*
- **Example**: Instead of just generating a reply, an agent might also **send it by email**, **schedule a meeting**, or **trigger a workflow**



Why AI Agents Matter

- Models create answers — **agents get things done**
- Agents are smart: they can **decide, adapt, and complete tasks**
- Let you go beyond content — **into automation and intelligent workflows**
- Useful for real-world enterprise applications

Use Cases for AI Agents - 1

- **Customer Agents**: Deliver 24/7 support across channels
  - *Example: A banking chatbot that handles account queries via app and voice*
- **Employee Agents**: Automate internal tasks and boost productivity
  - *Example: An HR agent that answers policy questions and translates internal memos*
- **Creative Agents**: Assist with content generation and design
  - *Example: A marketing agent that creates ad copy and campaign visuals*

Use Cases for AI Agents - 2

- **Data Agents**: Generate insights from complex datasets
  - *Example: An analytics agent that summarizes sales trends*
- **Code Agents**: Help developers write and review code
  - *Example: A coding assistant that autocompletes functions and explains APIs*
- **Security Agents**: Monitor and respond to security threats in real time
  - *Example: A threat detection agent that flags anomalies in login patterns*

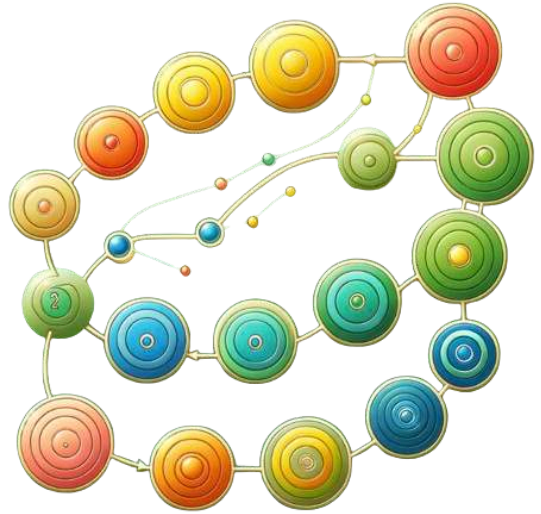
Agent Type: Conversational Agents

In 28 Minutes

- Understand natural language and respond intelligently
- **Flow**:
  - User sends a message
  - Agent identifies the **intent**
  - Fetches data or uses tools if needed
  - Crafts a response and replies
- **Examples**:
  - **Healthcare**: “Can I book an appointment with Dr. XYZ tomorrow?”
  - **Retail**: “Can you help me track my order?”

Agent Type: Workflow Agents

- **Workflow agents** automate multi-step business processes
- **Flow:**
  - Task is triggered (by user or system)
  - Agent plans and executes steps
  - Calls tools or systems
  - Delivers the result
- **Examples:**
  - **HR**: Onboard new employee – create email, assign equipment, schedule orientation
  - **IT Operations**: Restart crashed services, log incident, notify admin



Model vs Agent vs Application

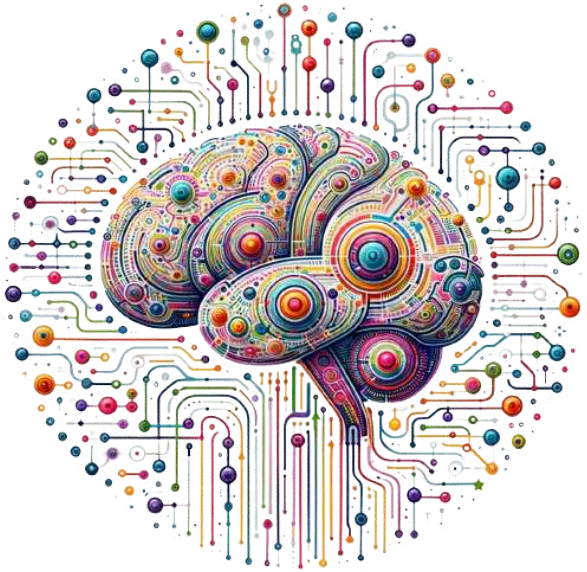
- **Model**: Core intelligence that generates content
  - Example: “Summarize this document”
- **Agent**: Goal-oriented system that uses models and tools to take action
  - Example: “Send summary to team and schedule a review”
- **Application**: User-facing interface that embeds one or more agents into a workflow
  - Example: Chatbot for managing leave requests

Multi-Agent Example: IT Support Assistant

- **Goal**: Streamline IT support with intelligent automation
- **Agent 1 - Troubleshooter**: Analyzes employee-reported problems using internal documentation and previous tickets
- **Agent 2 - Operations Manager**: Handles ticket creation, routing to the right team, and updating ticket statuses
- **Application UI**:
  - Conversational chat interface for reporting and tracking issues
  - Dashboard displaying ticket history, status, and resolution time
  - Real-time alerts and notifications for important updates
- **Key Benefit**: Reduces IT team's workload and accelerates issue resolution with minimal human intervention

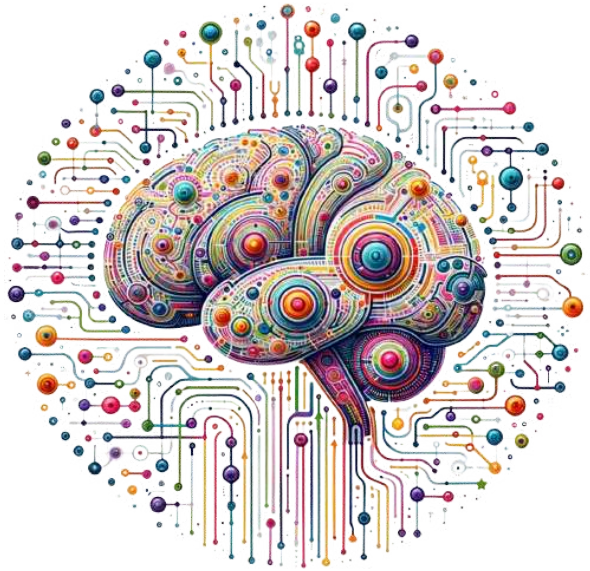
How Do AI Agents Work?

- Persona
- Model (LLM)
- Memory
- Tools



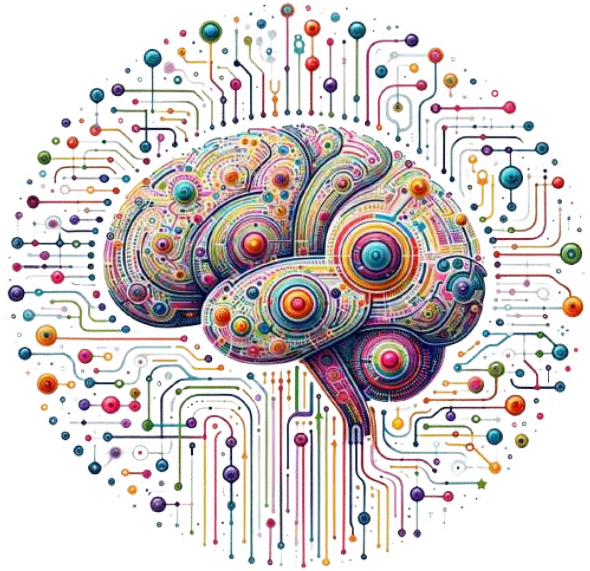
How Do AI Agents Work?

- **Persona**: Defines the agent's role, personality, and communication style
  - *Ensures consistent behavior aligned with the agent's purpose*
- **Model (LLM)**: The core reasoning engine for understanding and decision-making
  - Empowers agents to comprehend instructions, generate responses, and take action
- **Memory**: Enables learning and contextual awareness
  - Short-term: Handles immediate interaction
  - Long-term: Stores historical data and conversations



How Do AI Agents Work? - Tools

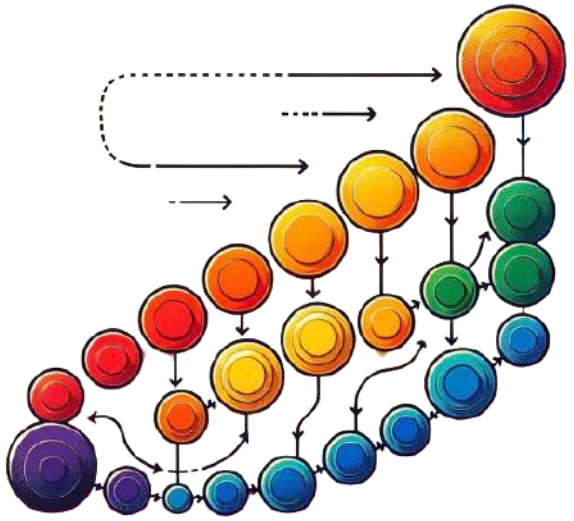
- **Tools**: Extend capabilities through interaction with external systems
  - **Extensions (APIs)**: Connect to external services
    - Ex: Translate API, Travel booking API
  - **Functions**: Built-in logic
    - Ex: "calculate_something()"
  - **Data Stores**: Pull or push data
    - *Ex: Customer DB, sales records*
  - **Plugins**: Add-ons like calendar or payment integration



Building Agents in Google Cloud

In 28 Minutes

- **AI Apps (Earlier Vertex AI Agent Builder)**
  - Create agents using natural language or code-first workflows
  - Ground agents in enterprise data using flexible integration options
- **Vertex AI Agent Engine**: Fully managed runtime for deploying your agents
- **Vertex AI Agent Garden (GitHub)**: Access a curated collection of pre-built agents and tools
  - Accelerate development with ready-made samples and frameworks



Understanding Need for Agentspace

In 28 Minutes

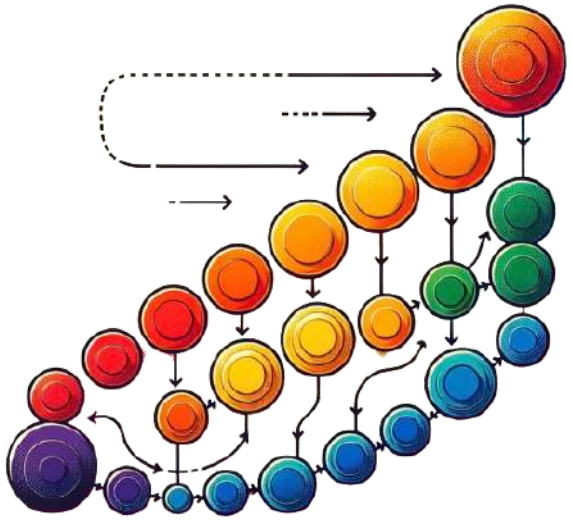
- Want to make your **employees more productive**?
- Want to make **enterprise information more discoverable**?
  - Even when spread across tools like Drive, Jira, Confluence, and SharePoint?
- **Google Agentspace** helps your team use your company's knowledge more effectively

What can you do with Agentspace?

- Create **custom AI agents** that access, understand, and act on data — wherever it's stored
- Integrate agents into **internal dashboards and websites**
- Acts as a **central launch point** for expert agents:
  - **NotebookLM Enterprise**: Upload docs, get insights, generate summaries—even audio recaps
  - **Multimodal Search Agents**: Search across structured and unstructured data from multiple systems
  - **Generative AI Assistants**: Use enterprise data to answer prompts and perform actions through connected tools

Vertex AI Search

- **Enterprise Search**: A fully managed, enterprise-grade search solution from Google Cloud
  - Unlock relevant, contextual search across your internal and external content
  - **Built On**: The same AI and infrastructure that powers Google Search
- **Supported Data Sources**: Websites, PDFs, documents, data from BigQuery, ..
- **Grounding**: Power generative AI apps with reliable knowledge:
  - Use as a RAG (Retrieval Augmented Generation) foundation for accurate, grounded responses



Generative AI Layers

Understanding the Layers of Generative AI

In 28 Minutes

- Generative AI has **multiple interconnected layers**
  - **Infrastructure**
  - **Model**
  - **Platform**
  - **Agent**
  - **Application**
- Each plays a role — from the **foundation (infrastructure)** to the **interface (application)**

Exploring Layers of Generative AI

- **Infrastructure**: The base computing layer
  - *Includes: GPUs, TPUs, cloud servers, storage, networking*
- **Model**: The core intelligence of Gen AI
  - *Examples: Gemini for text, Imagen for images,..*
- **Platform**: Tools to build, train, and deploy AI
  - *Ex: Vertex AI for model training and orchestration*
- **Agent**: Interacts with environment and takes actions
  - *Ex: Agents that handle specific tasks like order processing or report summaries*
- **Application**: User-facing AI experiences
  - *Ex: AI writing assistants, customer service bots, video generation tools*

Google's Customer Engagement Suite

Need for Customer Engagement Solutions

In 28 Minutes

- **Personalization**: Today's customers want quick answers and personal support, not just search results
- Typical enterprise support challenges:
  - Long wait times
  - Inconsistent service quality
  - Limited 24/7 support capabilities
- Every customer conversation is an opportunity for trust and retention
  - How can you modernize your customer interactions?
  - Enter **Google's Customer Engagement Suite**

Google's Customer Engagement Suite

- Many customers prefer **direct connection**, not just search
- Positive engagement can drive business success
- Google's Customer Engagement Suite offers:
  - **Conversational Agents**
  - **Agent Assist**
  - **Conversational Insights**
  - **Contact Center as a Service (CCaaS)**

Why Conversational Agents Matter

In28 Minutes

- Customers expect **immediate, natural, and accurate** support
- **Conversational agents**:
  - Leverage LLMs for human-like interaction
  - Support for text and voice-based virtual agents
  - Work across multiple channels: web, mobile, messaging platforms

Conversational Agents: Flexible and Powerful

- **Generative agents:**
  - Understand natural language, respond flexibly
  - Ideal for open-ended queries
- **Deterministic agents:**
  - Predefined flows (e.g., press 1 for sales)
  - Good for compliance and control
- **Hybrid agents:**
  - You can combine both as well

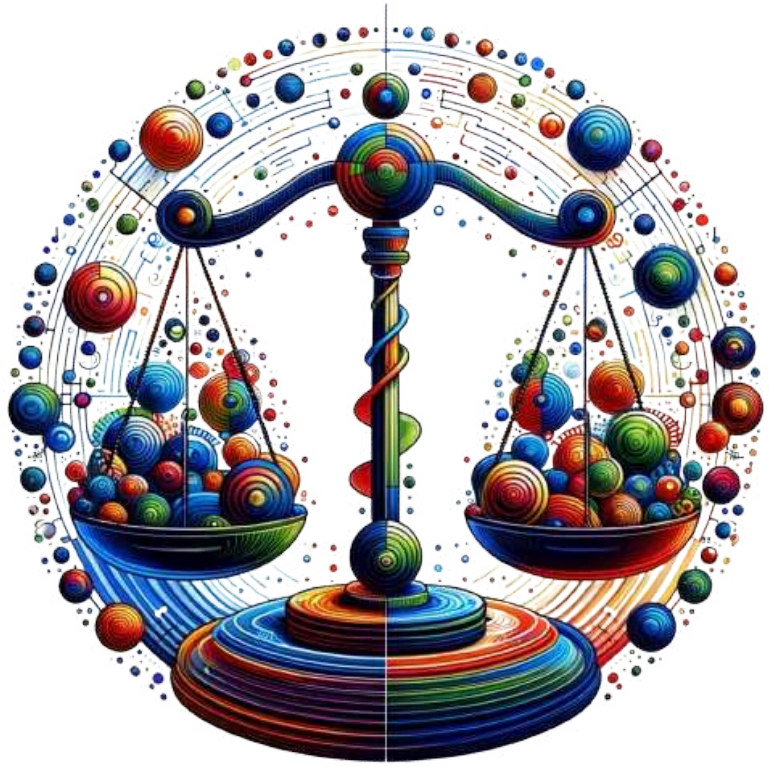
Agent Assist: Superpower for Live Agents

In28 Minutes

- **Live agents need real-time** help with:
  - Finding the right info
  - Handling multiple queries at once
  - Staying on brand
- Agent Assist **uses Generative AI** to:
  - Suggest context-aware responses
  - Pull up relevant documents or articles
  - Transcribe, translate, & summarize on the fly
- **Reduces errors** and improves customer satisfaction

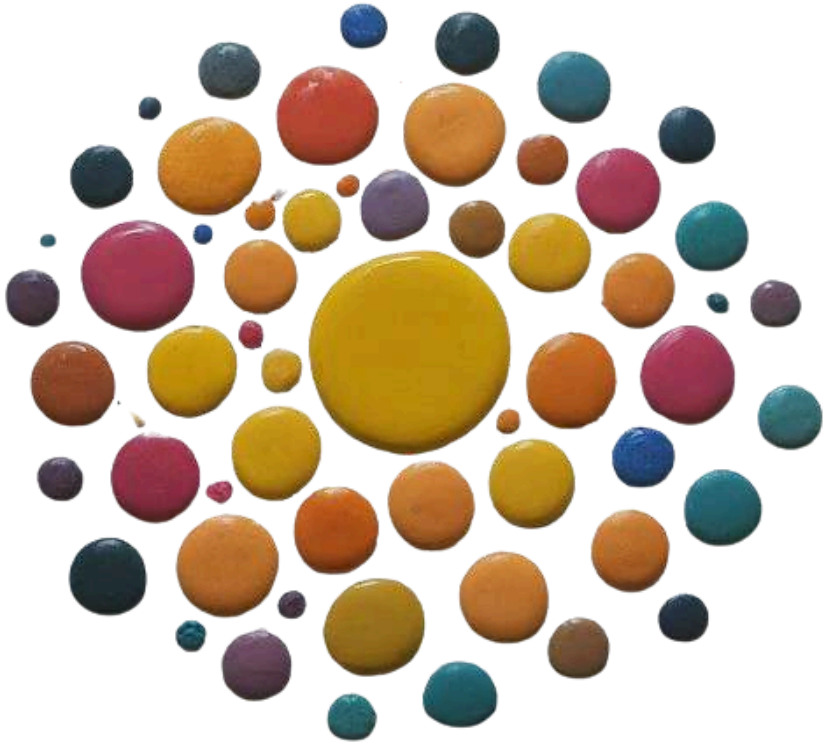
Conversational Insights: Learn from Every Interaction

- **Conversations** hold powerful business signals
- **Conversational Insights** uses ML to analyze interactions:
  - Detect emotions and sentiment trends
  - Identify common complaints or feature requests
  - Identify training needs for agents
- **Generate AI-powered FAQs** from chat and call transcripts
- **Drive continuous improvement** of your contact center



Contact Center as a Service (CCaaS): Modernized Support

- **Google's CCaaS** = Complete contact center platform in the cloud
- **Omnichannel by design**:
  - Supports chat, email, voice, video, etc.
  - Handles channel switching mid-conversation
- **Integrates** with CRM (e.g., Salesforce) and workforce tools
- Uses conversational agents, Agent Assist, and Insights



Customer Engagement Suite: End-to-End Benefits

In28 Minutes

- **Conversational Agents**: Automated chat using LLMs
- **Agent Assist**: Boosts live agent efficiency by enabling them
- **Conversational Insights**: AI-powered feedback loop
- **Contact Center as a Service (CCaaS)**: Cloud-native, scalable support hub
  - Combine all three!
  - Conversational Agents + Agent Assist + Conversational Insights

Scenarios for Google's Customer Engagement Suite

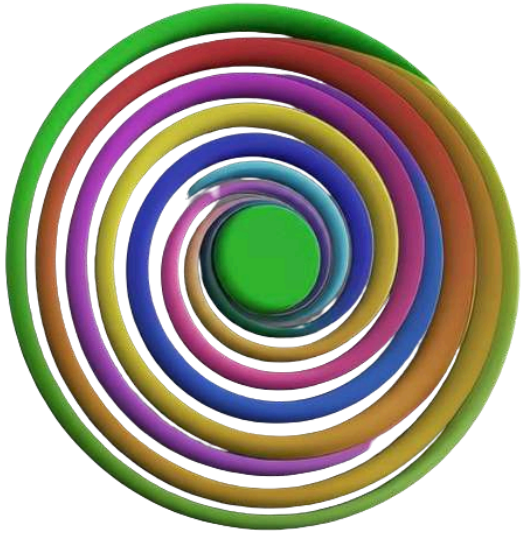
In 28 Minutes

Scenario	Solution
A retail brand wants to offer 24/7 product support via chat and voice	Conversational Agents
A telecom provider wants to help live agents by providing additional information (relevant documents or articles) to resolve issues faster	Agent Assist (Real-time suggestions and document lookup)
A bank wants to analyze customer calls to improve agent training	Conversational Insights
An airline wants to automate and setup a complete contact center platform in the cloud unifying support across chat, voice, and email	Contact Center as a Service (CCaaS)

Implementing Gen. AI In Your Organization

Strategic Plan for Gen AI Adoption

- **Establish a Clear Vision**: Align gen AI initiatives with your strategic business goals
- **Prioritize High-Impact Use Cases**: Focus on high-impact, measurable opportunities
- **Invest in Capabilities**: Build infrastructure, tools, and skills to support adoption
- **Drive Organizational Change**: Enable collaboration across business and technical teams
- **Champion Responsible AI**: Embed fairness, transparency, and security into all solutions

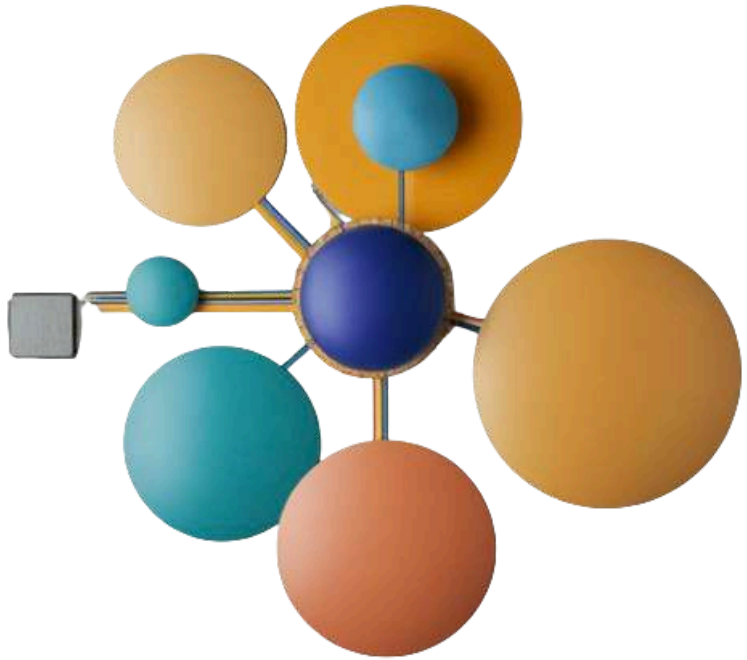


Measure Success and Manage Change

- **Define Key Metrics:**
  - ROI, revenue growth, efficiency, security, or customer experience
  - Measure financial benefits and operational improvements
- **Adapt to Change:**
  - Stay up-to-date with tools, models, and best practices
  - Encourage continuous learning through training and community engagement

The People: Key Roles in a Gen AI Project

- **Business Leaders:**
  - Identify high-impact use cases for AI
  - Help teams adopt AI in daily work
  - Ensure AI supports business goals and priorities
- **AI Practitioners:**
  - Choose and fine-tune custom Gen AI models
  - Ensure responsible AI practices
  - Handle scaling, data quality, and performance
- **Developers:**
  - Build AI-powered apps and assistants
  - Integrate Gen AI into existing systems and workflows
- *Success depends on cross-functional collaboration*



Who Does What in Gen AI Projects?

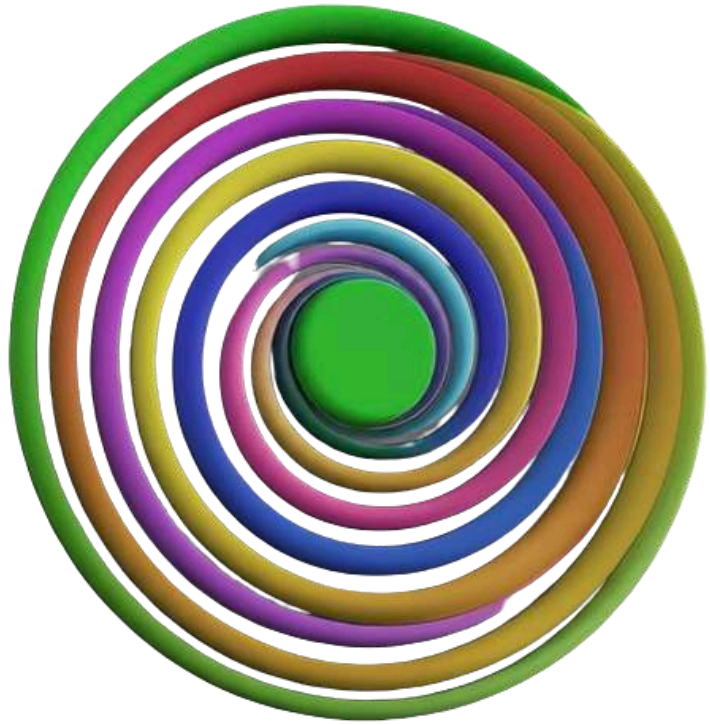
In 28 Minutes

Scenario	Role
Chooses the best Gen AI model and tunes it for company needs	AI Practitioner
Identify high-impact use cases for AI	Business Leader
Builds an internal chatbot that answers employee HR questions	Developer
Connects Gen AI output into an internal dashboard or CRM	Developer
Ensure AI supports business goals and priorities	Business Leader
Monitors model performance and retrains it with better data	AI Practitioner

Build a Gen AI Strategy: Combine Top-Down and Bottom-Up

In 28 Minutes

- Gen AI success needs **both a strategic vision and grassroots innovation**
- **Top-Down**:
  - Define a clear **Gen AI vision**
  - Identify and prioritize **high-value business use cases**
  - Allocate **funding, infrastructure, and talent**
- **Bottom-Up**:
  - Empower teams to **identify local problems and test solutions**
  - Leverage real-world feedback to refine strategy
- A **hybrid approach** drives **enterprise-wide alignment**



Top-Down: Executive-Led Gen AI Strategy - 1

In 28 Minutes

- **Strategic Focus**: Start with AI ideas that are high value (High-ROI) and easy to do (feasible)
  - **Example**: Automate document review using Gen AI
- **Exploration**: Support new ideas with events like hackathons and learning sessions
  - **Example**: Run a hackathon to build a tool that analyzes customer complaints
- **Responsible AI**: Set rules to make sure AI is safe, fair, and respects privacy
  - **Example**: Don't allow sharing of private data with public AI tools

Top-Down: Executive-Led Gen AI Strategy - 2

- **Resourcing**: Invest in platforms (e.g., Vertex AI), infrastructure, and skills
- **Impact**: Set goals (KPIs) and share how AI is helping
  - **Example**: Track how much money is saved after using AI to help with support calls
- **Continuous Improvement**: Review what's working, gather feedback, and adapt
  - **Example**: Hold monthly meetings to check progress and get inputs from stakeholders

Bottom-Up: Team-Driven Innovation

- **Strategic Focus**: Start with workflow challenges and user pain points
- **Exploration**: Try Gen AI tools; share wins and lessons
- **Responsible AI**: Align with company values; flag risks and test thoroughly
- **Resourcing**: Leverage available tools, and pilot environments; escalate needs to secure broader support
- **Impact**: Show value through metrics like time saved or satisfaction improved
- **Continuous Improvement**: Update models, prompts and processes based on feedback

Gen AI Strategy – Top-Down vs Bottom-Up Scenarios

In 28 Minutes

Scenario	Strategy
Leadership defines a goal to automate legal document review across regions	Top-Down
Support agents try Gen AI to summarize customer calls and reduce typing effort	Bottom-Up
Company launches an enterprise-wide AI vision and invests in Vertex AI	Top-Down
A small marketing team uses Gen AI to test headlines and social media copy	Bottom-Up
HR team experiments with Gen AI to auto-generate job descriptions	Bottom-Up
Executives define KPIs and review Gen AI progress every quarter	Top-Down

Using Gen AI: Augmentation vs Automation

- **Augmentation**: Gen AI supports complex thinking
  - Strategic planning → Gen AI trend analysis + human leadership
  - Creativity → Gen AI idea generation + human innovation
  - Problem solving → Gen AI insights + human judgment
- **Automation**: Gen AI handles repetitive tasks
  - Data entry, content formatting, basic summarization
- AI **frees humans** to do higher-value work

Generative AI – Augmentation vs Automation Scenarios

Scenario	Type
Gen AI answers customer FAQs on your website without human help	Automation
Gen AI creates draft email responses; user reviews and sends	Augmentation
Gen AI summarizes long documents for legal teams to review	Augmentation
Gen AI translates support tickets into English and auto-replies with template	Automation
Gen AI generates reports and charts that a manager uses to present strategy	Augmentation
Gen AI schedules meetings automatically based on calendar availability	Automation

Keeping Humans at the Forefront of Gen AI

- Gen AI is powerful, but **humans must stay in control**
- Use Gen AI but ensure people are:
  - Making decisions
  - Interpreting outputs
  - Providing feedback
- The best results come from **human-AI collaboration**

Humans in the Loop (HITL)

- ML models are powerful, but some tasks need **human judgment and oversight**
- **HITL** integrates human feedback before, during, or after model use
  - **Content Moderation**: Filter subtle content AI might miss
    - *Ex: Flagging sarcastic hate speech in social media comments*
  - **High-Risk Decisions**: Ensure accountability in critical systems
    - *Ex: Validating AI-based loan approval decisions*
  - **Post-Generation Review**: Continuous improvement
    - *Ex: Reviewing customer support chatbot chats for accuracy and tone*

Secure AI

Why Secure AI at Every Stage?

- AI systems are only as secure as their weakest point
- Security must span from data ingestion to post-deployment management
- Plan. Secure. Monitor. Improve. Repeat.
- Let’s explore how to secure each stage in the AI lifecycle

Secure AI: At Each Stage

- Secure AI is important at each stage
- **1:** Data Ingestion
- **2:** Data Preparation
- **3:** Model Training
- **4:** Model Deployment
- **5:** Model Management

Secure AI: Protecting AI From End to End - 2

In28 Minutes

Stage	Key Security Actions
Data Ingestion	Restrict data upload/edit access. Accept only trusted sources. Log all ingestion activity.
Data Preparation	Anonymize sensitive info. Encrypt data. Validate inputs before using for model training.
Model Training	Secure training environment.Restrict access to model configuration. Prevent model theft.
Model Deployment	Control who can query the model. Secure APIs with authorization & rate limits. Monitor usage.
Model Management	Monitor for drift and bias. Use alerting tools.

Secure AI – Match the Action to the Stage

Action	Stage
Log every file that gets uploaded to the training folder	Data Ingestion
Prevent unauthorized access to a pricing model's API	Model Deployment
Alert team when prediction accuracy suddenly drops	Model Management
Train models in an isolated environment to avoid external access	Model Training
Replace email addresses with masked IDs before feeding into model	Data Preparation
Detect bias over time and update model configuration	Model Management

Secure AI: Tools and Frameworks

- **Secure AI Framework (SAIF)** by Google
  - Designed to **detect threats**, **automate protection**, and **manage risk**
  - Integrates with existing enterprise security systems
- **Google Cloud Security Tools**
  - **Secure-by-design Infrastructure**: Private network + encryption
  - **IAM**: Control who can access what
  - **Security Command Center**: Central view of cloud security posture
  - **Observability**: Monitor your cloud deployments

Responsible AI

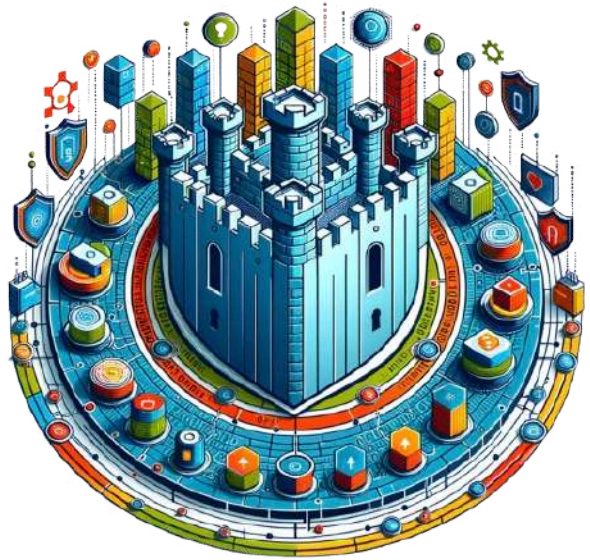
Responsible AI: Build Trustworthy AI Systems

In 28 Minutes

- **Responsible AI** ensures that applications avoid intentional and unintentional harm
- It's about developing AI ethically with **security**, **fairness**, and **transparency**
- Responsible AI = **safe, fair, transparent, and accountable AI**

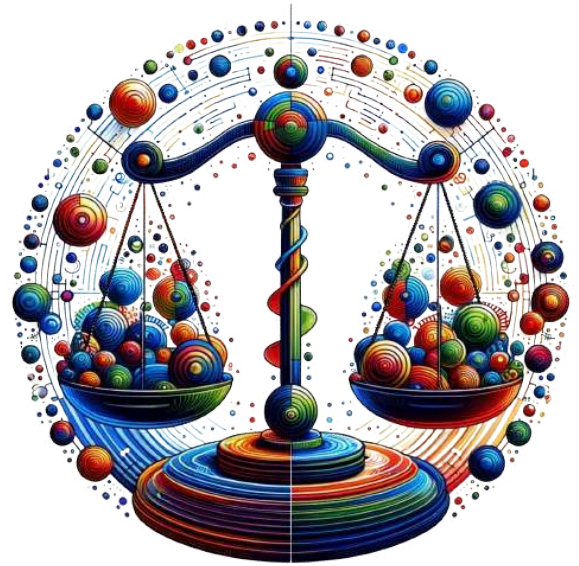
Secure Foundations for Responsible AI

- **Security is the base layer** of all responsible AI practices
  - Prevent data theft, misuse, and manipulation
  - Protect from malicious attacks
- *Ex: Encrypting customer data used for training a recommendation engine*
- Without security, transparency, privacy, and fairness **cannot be guaranteed**



Transparency & Privacy

- **Transparency**: Users should understand how AI makes decisions
  - *Ex: Disclosing how a loan approval model evaluates applications*
  - Explainability: Explaining the reasoning behind the AI's decisions
- **Privacy**: Safeguard sensitive user data
  - *Ex: Anonymizing data before training a model to prevent leakage*
  - Implement safeguards to prevent models from exposing training data



Accountability, Explainability & Legal Compliance

In 28 Minutes

- **Accountability**: Someone must be responsible for AI's decisions
- **Explainability**: Understand how the AI reached a conclusion
  - *Ex: Vertex Explainable AI helps visualize which features influenced a prediction*
- **Legal Compliance**: AI systems must comply with various laws - privacy, IP, anti-discrimination,..
  - *Ex: GDPR requires transparency in algorithmic decision-making*
  - Know your model's licensing terms and data use policies

Obtaining Right Data: Data Accessibility

In 28 Minutes

- **Availability**: Data may exist but be locked away
  - *Ex: Health data hidden due to privacy regulations*
- **Cost**: High-quality data isn't free
  - *Ex: Labeling support emails can cost time and money*
- **Format**: Raw data may need cleanup
  - *Ex: Handwritten forms must be digitized*
- **Usability**: Noisy or messy data reduces model effectiveness
  - *Ex: Social media slang must be cleaned before analysis*

Ensuring Data Quality

- **Accuracy**: Wrong labels = wrong learning
  - *Ex: A cat labeled as a "dog" confuses the model*
- **Completeness**: Too little data = weak insights
  - *Ex: One-day weather history isn't enough for forecasting*
- **Representativeness**: Missing groups = biased results
  - *Ex: Ignoring teens in shopping data skews recommendations*
- **Consistency**: Mixed formats = model confusion
  - *Ex: "NYC" vs "New York" vs "NewYork"*

Fairness

- **Bad data = bad outcomes**
  - Poor, incomplete, or biased data skews outcomes and can cause real-world harm
- **Bias in Data**: Can lead to unfair predictions
  - *Ex: Loan models denying minority applicants*
- **Fairness**: How do we ensure models treat all users fairly?
  - **Audit your data**: Check for underrepresented groups
  - **Balance datasets**: Include diverse examples
  - **Use fairness tools**: Try What-If Tool or Fairness Indicators
  - **Test across groups**: Validate model accuracy across age, gender, region

Get Ready

Generative AI Leader - Certification Resources

In28 Minutes

Title	Link
Home Page	*https://cloud.google.com/learn/certification/generative-ai-leader*
Exam Guide	Link on Home Page
Sample Questions	Link on Home Page
Registering For Exam	Link on Home Page

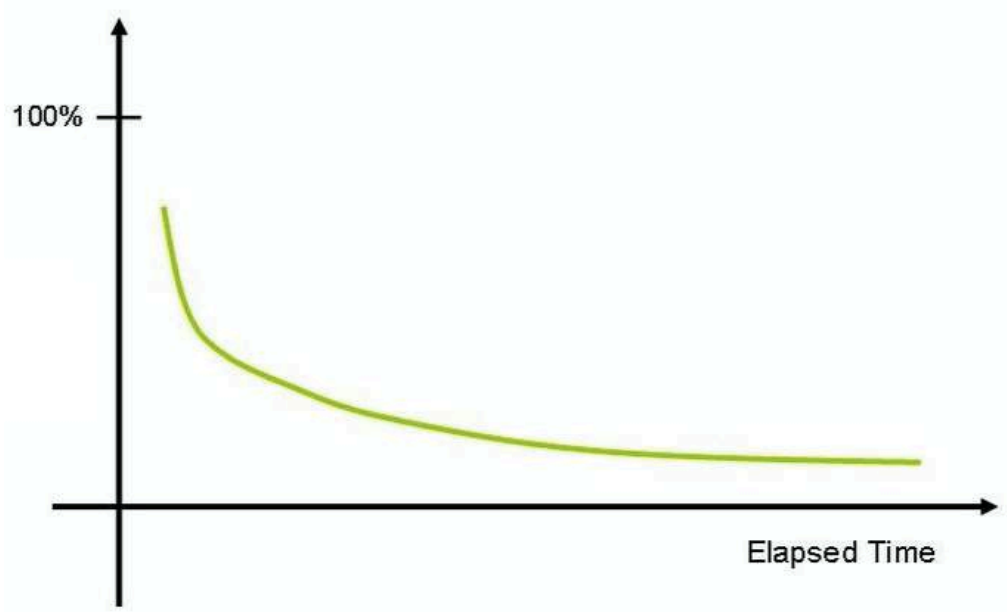
Generative AI Leader - Certification Exam

In28 Minutes

- **50-60 multiple choice questions** and **90 Minutes**
  - **No penalty** for wrong answers
  - **Questions**:
    - Multiple Choice - 4 options and 1 right answer
  - Result immediately shown after exam completion
    - Email (a few days later)
- My Recommendations:
  - Read the **entire question**
    - Identify and write down the **key parts of the question**
  - More than sufficient time
  - **Flag questions** for future consideration (Review before final submission)
  - TIP: **Answer by Elimination!**

Get Ready For Your Exam

- How do you improve your chances of remembering things for the exam?
  - **1:** Review the presentation
  - **2:** Watch videos again at 2X speed



You are all set!

Let's clap for you!

- You have a lot of patience! **Congratulations**
- You have put your best foot forward to be an Google Generative AI Leader
- Make sure you prepare well
- Good Luck!

Do Not Forget!

- Recommend the course to your friends!
  - **Do not forget to review**!
- **Your Success = My Success**
  - Share your success story with us on LinkedIn (Tag - in28minutes)
  - Share your success story and lessons learnt in Q&A with other learners!

What Next?

FASTEST ROADMAPS

in28minutes.com

In28 Minutes

- Google Cloud Certifications
- Azure Certifications
- AWS Certifications
- DevOps
- Java Full Stack
- Java Microservices